TAS-EGNN: Task-Aware Spectral Ego-Graph Neural Network for Scalable Graph Coreset Selection

Mebarka Allaoui[1] Rachid Hedjam[1] Sonia Gupta[2]
[1] Bishop's University, Sherbrooke, QC, Canada [2] AI Garage, Mastercard, Gurugram, India

Why This Work Matters

- Training GNNs on large graphs is expensive in memory and runtime.
- Classical subset selection methods ignore graph structure or task feedback.
- Condensation/coarsening methods are often too heavy to scale.
- **Goal:** select compact graph coresets that preserve performance while remaining practical on large real-world graphs.

TAS-EGNN: Core Idea

TAS-EGNN builds a compact yet informative graph coreset by scoring each training node with a **task-aware local criterion** and then selecting a **diverse set of ego-graphs** under a budget. The score combines complementary signals from structure, prediction uncertainty, and supervised error:

$$s(v) = \alpha\,\mathcal{H}_{\mathrm{spec}}(\mathcal{E}_v) + \beta\,\mathcal{H}_{\mathrm{pred}}(v) + \gamma\,\mathbb{1}[\hat{y}_v \neq y_v].$$

- **Structure-aware:** spectral entropy favors heterogeneous and informative local neighborhoods.

$$\mathcal{H}_{\mathrm{spec}}(\mathcal{E}_v) = -\sum_{i=1}^{k} \tilde{\lambda}_i \log \tilde{\lambda}_i$$

- **Task-aware:** predictive entropy and misclassification emphasize ambiguous and hard training nodes.

$$\mathcal{H}_{\mathrm{pred}}(v) = -\sum_{c=1}^{C} p_v^{(c)} \log p_v^{(c)}, \qquad \mathbb{1}[\hat{y}_v \neq y_v]$$

- **Diversity-aware:** greedy coverage penalizes overlapping ego-graphs, reducing redundancy in the selected subset.

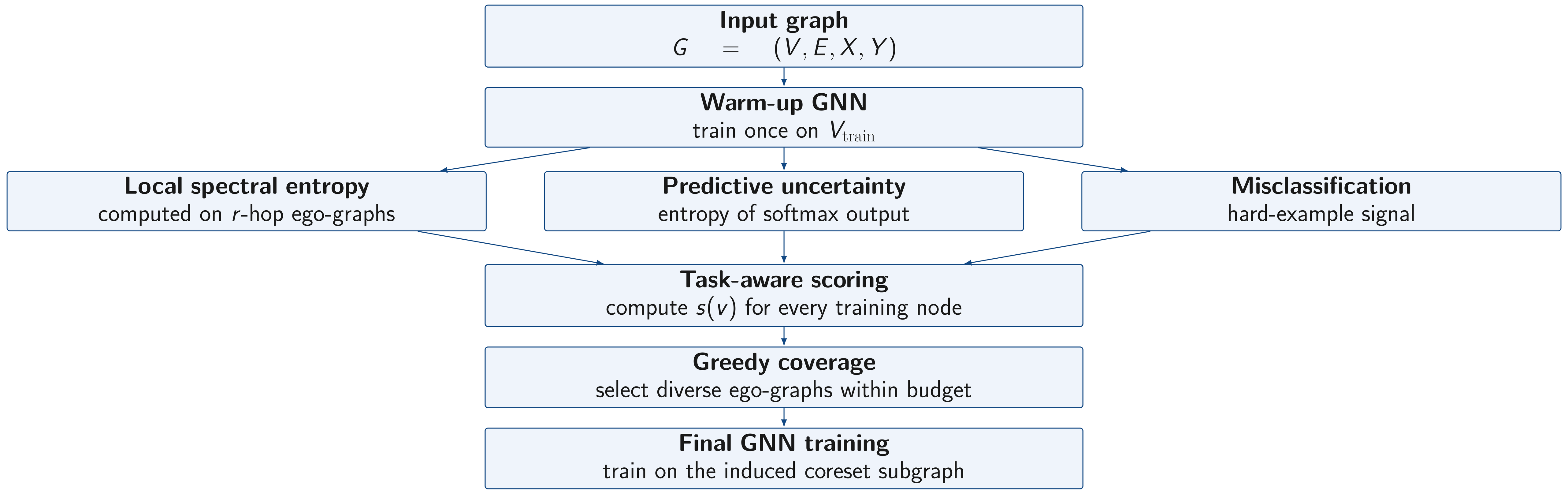
$$v^* = \arg\max_{v \in V_{\mathrm{train}} \setminus \mathcal{S}} \Big[s(v) - \tau \sum_{u \in \mathcal{S}} \mathrm{Overlap}(\mathcal{E}_u, \mathcal{E}_v) \Big]$$

- **Outcome:** a scalable coreset strategy that preserves strong GNN performance on citation, social, fraud, and industrial-scale graphs.

Why it Scales and Generalizes

- **Localized computation:** operates on ego-graphs, avoiding expensive full-graph spectral analysis.
- **Reduced redundancy:** greedy coverage discourages overlapping ego-graphs and improves subset diversity.
- **Backbone-agnostic:** integrates well with GCN, GraphSAGE, GAT, GIN, and APPNP [2, 6].
- **Robust in practice:** remains effective across datasets of very different scales and domains.

Model Overview



TAS-EGNN's Algorithm

Require: Graph $\mathcal{G} = (V, E, X, Y)$, GNN f_θ, scoring weights α, β, γ, coreset budget c
Ensure: Selected coreset $\mathcal{S}$

Initialize $\mathcal{S} \leftarrow \emptyset$
Compute predictions $\hat{Y} = f_\theta(X, A)$ and softmax probabilities p_v for all $v \in V_{\mathrm{train}}$
for each node $v \in V_{\mathrm{train}}$ **do**
  $\mathcal{E}_v \leftarrow r$-hop ego-graph
  $s(v) \leftarrow \alpha\,\mathcal{H}_{\mathrm{spec}}(\mathcal{E}_v) + \beta\,\mathcal{H}_{\mathrm{pred}}(v) + \gamma\,\mathbb{1}[\hat{y}_v \neq y_v]$
end for
while $|\mathcal{S}| < c$ **do**
  $v^* \leftarrow \arg\max_{v \in V_{\mathrm{train}} \setminus \mathcal{S}} [s(v) - \tau \sum_{u \in \mathcal{S}} \mathrm{Overlap}(\mathcal{E}_u, \mathcal{E}_v)]$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{v^*\}$
end while
return $\mathcal{S}$

Datasets

Dataset	Nodes	Edges	Feats	Classes
Cora	2,708	5,429	1,433	7
Citeseer	3,327	4,732	3,703	6
Flickr	89,250	899,756	500	7
OGBN-Arxiv	169,343	1,166,243	128	40
OGBN-Products	2,449,029	123,718,152	100	47
Banksim	17,200	98,879	7	2
Paysim	18,213	100,654	7	2
ECC	20,000	108,800	30	2

Backbone Robustness

Dataset	Metric	GCN	GraphSAGE
Citeseer	Accuracy	77.1±0.4	**80.3±0.5**
Cora	Accuracy	86.8±0.1	**92.3±0.2**
Flickr	Accuracy	59.7±0.2	**75.8±1.6**
OGBN-Arxiv	Accuracy	65.1±0.2	**68.6±0.1**
OGBN-Products	Accuracy	65.2±0.4	**69.5±0.6**
Banksim	PR-AUC	93.1±1.5	**96.5±0.5**
Paysim	PR-AUC	95.6±0.9	**96.1±1.4**
ECC	PR-AUC	93.8±0.6	**97.6±0.2**

Main Results

Dataset	Ratio	Best competitor	TAS-EGNN
Citeseer	50%	75.3	**80.3**
Cora	50%	87.7	**92.3**
Flickr	2.0%	64.4	**75.8**
OGBN-Arxiv	2.0%	67.6	**68.6**
OGBN-Products	2.0%	45.0	**69.5**
Banksim	50%	94.2±0.1	**96.5**
Paysim	50%	93.3±0.0	**96.1**
ECC	50%	91.6±0.2	**97.6**

Ablation Summary

Variant	Citeseer	Flickr	Paysim	ECC
$s(v)$	**80.3±0.5**	**75.8±1.6**	**96.1±1.4**	**97.6±0.2**
$\alpha \cdot \mathcal{H}_{\mathrm{spec}}(\mathcal{E}_v) + \beta \cdot \mathcal{H}_{\mathrm{pred}}(v)$	80.2±0.4	74.8±0.8	96.0±1.2	97.2±0.5
$\alpha \cdot \mathcal{H}_{\mathrm{spec}}(\mathcal{E}_v) + \gamma \cdot \mathbb{1}[\hat{y}_v \neq y_v]$	79.0±0.3	75.3±1.4	95.8±1.5	97.4±0.3
$\beta \cdot \mathcal{H}_{\mathrm{pred}}(v) + \gamma \cdot \mathbb{1}[\hat{y}_v \neq y_v]$	80.1±0.2	75.5±0.6	96.0±1.3	97.0±0.3
$\mathcal{H}_{\mathrm{spec}}(\mathcal{E}_v)$	78.9±0.2	74.1±1.8	95.1±1.5	94.4±2.4
$\mathcal{H}_{\mathrm{pred}}(v)$	79.1±0.1	75.4±0.7	95.7±0.6	95.1±0.5
$\mathbb{1}[\hat{y}_v \neq y_v]$	75.0±0.4	75.0±1.1	95.7±1.1	94.2±1.6

Take-Home Message

TAS-EGNN turns **local structure**, **predictive uncertainty**, and **hard-example feedback** into a compact and diverse training subset, enabling **strong accuracy** and **practical scalability** on real graphs.

Practical Efficiency

- **Memory-efficient:** lower memory usage than graph condensation and coarsening methods [1, 4, 5].
- **Lightweight:** no full-graph eigendecomposition and no synthetic graph generation.
- **Deployable:** effective from small benchmarks to very large OGB graphs [3].

References

[1] Xinyi Gao, Guanhua Ye, Tong Chen, Wentao Zhang, Junliang Yu, and Hongzhi Yin. Rethinking and accelerating graph condensation: A training-free approach with class partition. In *Proceedings of the ACM on Web Conference 2025*, pages 4359–4373, 2025.

[2] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[3] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

[4] Wei Jin, Lingxiao Zhao, Shichang Zhang, Yozen Liu, Jiliang Tang, and Neil Shah. Graph condensation for graph neural networks. *arXiv preprint arXiv:2110.07580*, 2021.

[5] Mohit Kataria, Sandeep Kumar, et al. Ugc: Universal graph coarsening. *Advances in Neural Information Processing Systems*, 37:63057–63081, 2024.

[6] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Paper

GitHub Repository

Contact: mallaoui@ubishops.ca