Partition-Based Active Learning for Graph Neural Networks

Jiaqi Ma*, Ziqiao Ma*, Joyce Chai, Qiaozhu Mei

University of Michigan FORESEER Group

Content · Context · Crowd · Cloud

Summary

- We study semi-supervised learning with Graph Neural Networks (GNNs) in an one-shot Active Learning (AL) setup.
- We propose GraphPart, a novel partition-based active learning approach for GNNs without additional hyperparameters.
- Extensive experiments on multiple benchmark datasets demonstrate that GraphPart outperforms existing under a wide range of annotation budget constraints.

Motivations

We study GSSL in an one-shot AL setup:

- Realistic Setting: We have access to abundant unlabeled samples prior to learning, and flexibility to query labels for a small portion of the samples.
- Framework Nature: GNNs utilize the relational information among the interconnected samples, and properly selecting nodes to annotate may further enhance GNN's performance.

Limitations: Prior methods were unable to fully utilize the smoothness properties of graph, including *local smoothness* and *global smoothness*.

Method

The GraphPart approximates a optimization problem that is equivalent to minimizing the objective of a K-Medoids problem with on each graph partition.

Algorithm 1 Graph-Partition-Based Query

Input: A K-partition $\mathcal{T}_K$ of the graph, budget b

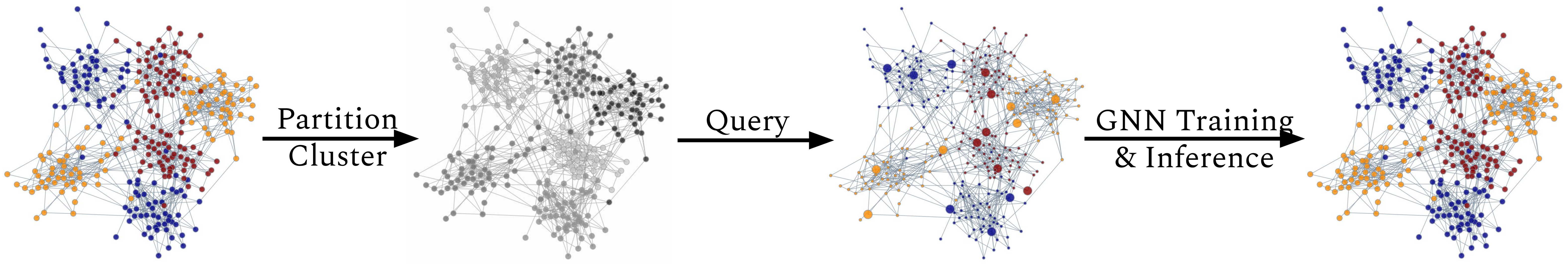
Output: A subset of unlabelled nodes $\mathbf{s}_1$ of size b: $(\mathbf{s}_1 \subseteq V \setminus \mathbf{s}_0$ and $|\mathbf{s}_1| = b)$

- Set $\mathbf{s}_1 = \emptyset$.
- **for** $T_k \in \mathcal{T}_K$ **do**
  - $b_k \leftarrow b//K$.
  - $T_k \leftarrow T_k \setminus \{\mathbf{s}_0 \cup \mathbf{s}_1\}$.
  - $E_k \leftarrow \{g(v_i)\}_{i \in T_k}$.
  - $\mathbf{s} \leftarrow b_k$-Medoids$(E_k)$.
    //Perform K-Medoids clustering on the set of data points E_k with b_k medoids returned as $\mathbf{s}$.
  - $\mathbf{s}_1 = \mathbf{s}_1 \cup \mathbf{s}$.
- **end for**
- **return** $\mathbf{s}_1$

We also introduce GraphPartFar, a greedy correction that instead of selecting nodes closest to centers, the distance function to minimize is penalized by the minimum distance to any selected node. GraphPartFar makes sure that all the selected nodes are not too close and similar to each other, increasing the diversity of the pool.

Formulation

Combining the settings of one-shot learning and batch-mode active learning: In each run, the algorithm use up the predefined budget to select a batch of nodes to label. The querying process is done once and for all in order to minimize retraining.



Theoretic Analysis

Assumption 1 (Label Smoothness). *Assume that* $\forall c \in [C]$, *there exists a function* $\eta_c : \mathcal{V} \to [0, 1]$ *such that* $\Pr[y_i = c \mid v_i] = \eta_c(v_i)$ *for any* $i \in V$. *Moreover,* $\forall k \in [K]$, $\forall i, j \in T_k$, *assume that there exists a constant* $\delta_\eta < \infty$, *such that*

$$|\eta_c(v_i) - \eta_c(v_j)| \le \delta_\eta ||g(v_i) - g(v_j)||_2.$$

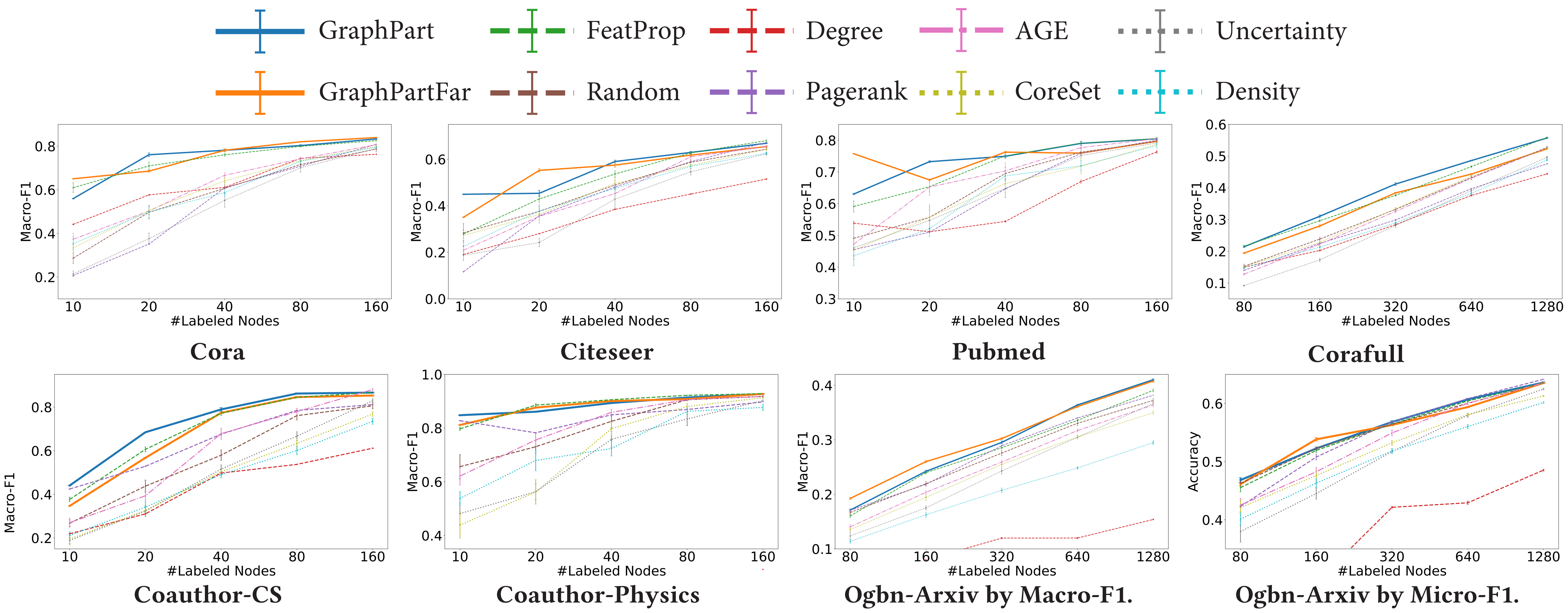
Assumption 2 (Model Smoothness). *Assume that* $\forall e, e' \in \mathbb{R}^{d'}$, *the MLP* h *satisfies* $||h(e) - h(e')||_\infty \le \delta_h ||e - e'||_2$ *for some constant* $\delta_h < \infty$.

The Main Result. We use $T(i)$ to denote the partition where the node i belongs to, and denote for convenience the training set $S_{tr} := \mathbf{s}_0 \cup \mathbf{s}_1$ and the test set $S_{te} := V \setminus S_{tr}$. We have:

Proposition 1. *For any fixed GNN model* f, *under Assump-tions 1 and 2, for any* $i \in S_{te}$, *if* $S_{tr} \cap T(i) \neq \emptyset$, *letting* $\tau(i) := \arg\min_{l \in S_{tr} \cap T(i)} ||g(v_i) - g(v_l)||_2$, $\epsilon_i := ||g(v_i) - g(v_{\tau(i)})||_2$, *and* $\gamma_i := 2\delta_h \epsilon_i$, *then we have*

$$\mathbb{E}_{y_i}[\mathcal{L}_0(f(v_i), y_i)] \le C\delta_\eta \epsilon_i + \mathbb{E}_{y_{\tau(i)}}[\mathcal{L}_{\gamma_i}(f(v_{\tau(i)}), y_{\tau(i)})].$$

Experiments



Bias Mitigation