An Instance-Level Profiling Framework for Graph-Structured Data

Tianqi Zhao[1], Russa Biswas[2], Megha Khosla[1]

[1] TU Delft
[2] Aalborg University

Graph machine learning models often achieve similar overall performance yet behave differently at the node level—failing on different subsets of nodes with varying reliability. Standard evaluation metrics such as accuracy obscure these fine-grained differences, making it difficult to diagnose when and where models fail. We introduce NODEPRO, a node profiling framework that combines *data-centric* signals capturing feature dissimilarity, label uncertainty, and structural ambiguity with *model-centric* measures of prediction confidence and consistency to provide fine-grained insights into node-level behavior and failure modes.

Our Proposed Framework. We introduce **NODEPRO**, that integrates *data-centric* and *model-centric* profiling scores to characterize node difficulty, uncertainty, and prediction reliability beyond aggregate metrics. The two profiling scores are explained below. We use the following notations. Given a graph $\mathscr{G} = \{\mathscr{V}, \mathscr{E}\}$, each node $v \in \mathscr{V}$ has a feature vector $\mathbf{x}_v \in \mathbb{R}^d$ and a label $y_v \in \{1, \ldots, C\}$. The one-hot label is $\mathbf{y}_v \in \{0,1\}^C$, and $\mathscr{V}_c$ denotes nodes in class c.

Data-Centric Node Profiling. We define three types of scores that quantify the intrinsic difficulty of a node based on its input features, neighborhood consistency, and higher-order structural context. Firstly, we define the **Intra-class Feature Dissimilarity (ICFD)** which measures how atypical a node's features are compared to others in its class by computing the average cosine dissimilarity between a node's feature vector $\mathbf{x}_v$ and those of its class peers $\mathscr{V}c$: $S_{f_v} = 1 - \frac{1}{|\mathscr{V}c|-1} \sum v' \in \mathscr{V}c \setminus v \frac{\mathbf{x}_v \cdot \mathbf{x}_{v'}}{|\mathbf{x}_v||\mathbf{x}_{v'}|}$. A higher S_{f_v} indicates that the node has atypical features relative to its class, making it harder to classify using feature-based models, whereas a lower value implies stronger intra-class alignment.

Secondly, **Neighborhood Class Divergence (NCD)** evaluates how consistent a node's one-hop neighborhood is with those of other nodes in the same class. Let $\mathscr{P}_v(c)$ be the normalized label distribution in the neighborhood $\mathscr{N}(v)$, $\mathscr{P}_v(c) = \frac{\sum_{u \in \mathscr{N}(v)} \mathbf{y}_{u,c}}{\sum_{u \in \mathscr{N}(v)} \sum_{c'=1}^{C} \mathbf{y}_{u,c'}}$, and $\mathscr{Q}_{y_v}(c)$ the average distribution over neighborhoods of nodes in class c, $\mathscr{Q}_{y_v}(c) = \frac{1}{|\mathscr{V}_c|} \sum_{v' \in \mathscr{V}_c} \sum_{u \in \mathscr{N}(v')} \mathbf{y}_{u,c}$. We compute the divergence as $S_{l_v} = \sum_{c \in \mathscr{C}} \big[\log(\mathscr{P}_v(c) + \varepsilon) - \log(\mathscr{Q}_{y_v}(c))\big](\mathscr{P}_v(c) + \varepsilon)$, where $\varepsilon = 10^{-10}$ ensures numerical stability. A higher S_{l_v} indicates greater neighborhood inconsistency.

Lastly, the **Random Walk Class Divergence (RWCD)** captures higher-order label mixing around a node by analyzing the class distribution encountered during random walks. For node v, let $\mathscr{S}$ be the multiset of nodes visited across N walks of length k, and define the label count vector $\mathbf{d}w = \sum j \in \mathscr{S} \mathbf{y}_j$. The divergence $S_h = 1 - \frac{\mathbf{d}w[c]}{\sum c' \in \mathscr{C} \mathbf{d}_w[c']}$

quantifies how mixed the surrounding neighborhood is, where a higher S_h indicates more heterogeneous neighborhoods and thus greater classification difficulty.

Model-Centric Node Profiling. Following pior work Seedat et al. 2022, given a model $\mathscr{M}_{(\theta)}$, we trained on $\mathscr{G}$ with checkpoints $\mathscr{E} = \{e_1, \ldots, e_E\}$, let $\mathscr{P}_c(v, \theta_e)$ denote the predicted probability of node v's true class c at checkpoint e. The mean predicted probability is $\overline{\mathscr{P}}(v) = \frac{1}{E}\sum_{e\in\mathscr{E}} \mathscr{P}_c(v, \theta_e)$. Two complementary uncertainties are computed:

$$v_{\text{ep}}(v) = \frac{1}{E}\sum_{e}\big(\mathscr{P}_c(v,\theta_e) - \overline{\mathscr{P}}(v)\big)^2, \quad v_{\text{al}}(v) = \frac{1}{E}\sum_{e}\mathscr{P}_c(v,\theta_e)\big(1-\mathscr{P}_c(v,\theta_e)\big).$$

Epistemic uncertainty (v_{ep}) reflects prediction instability, while aleatoric uncertainty (v_{al}) captures intrinsic label ambiguity. Nodes are categorized as

$$g(v,\mathscr{G}) = \begin{cases} \text{Easy} & \overline{\mathscr{P}}(v) \geq C_{\text{up}} \wedge v_{al}(v) < P_{50}[v_{al}], \\ \text{Hard} & \overline{\mathscr{P}}(v) \leq C_{\text{low}} \wedge v_{al}(v) < P_{50}[v_{al}], \\ \text{Ambiguous} & \text{otherwise.} \end{cases}$$

Key Results

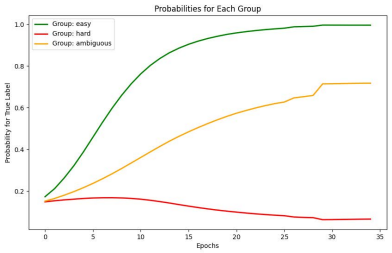

Fig. 1: GCN trained on CORA.

	CORA	CREDIT	BITCOINALPHA
GCN	0.749	0.679	0.912
GAT	0.793	0.619	0.846
GRAPHSAGE	0.839	0.630	0.722
MLP	0.752	0.587	0.951

Table 1: Accuracy of difficulty categorization of test nodes.

Model behavior analysis. We apply NODEPRO to compare different models in terms of their learning behavior. As shown in Figure 1, the node categorized by NODEPRO shows clearly different training patterns during training. One could then ask on how do these different node classifications relate to their data centric scores? What are the properties of nodes over which a model finds hard to learn from? What is the impact of such hardness on its generalization behavior?

Inductive Profiling of Unseen Nodes. We estimate profile of a new node with NODEPRO as follows. For a new node v_{new} with features $\mathbf{x}_{\text{new}}$ and edges $\mathscr{E}_{\text{new}}$, we compute its embedding using the final checkpoint. Its difficulty label (easy/hard/ambiguous) is the majority category among its K-nearest neighbors in embedding space. The obtained label is compared against the model centric profile obtained using its true label to obtain accuracy in categorization as reported in Table 1.

References

Seedat, Nabeel et al. (2022). *Data-IQ: Characterizing subgroups with heterogeneous outcomes in tabular data*. arXiv: 2210.13043 [cs.LG]. URL: https://arxiv.org/abs/2210.13043.