Proof: Incorporating Node Features Reduces Jensen-Shannon Divergence

To prove that incorporating node features reduces the Jensen-Shannon Divergence (JS divergence) $D_{\text{JS}}(P(G_s), P(G))$, we proceed as follows:

1. Jensen-Shannon Divergence Definition

The JS divergence between two probability distributions P and Q is defined as:

$$D_{\text{JS}}(P\|Q) = \frac{1}{2}D_{\text{KL}}(P\|M) + \frac{1}{2}D_{\text{KL}}(Q\|M),$$

where $M = \frac{1}{2}(P+Q)$ is the midpoint (average) distribution, and $D_{\text{KL}}(P\|Q)$ is the Kullback-Leibler (KL) divergence:

$$D_{\text{KL}}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

In this case:

- $P(G_s)$: Distribution of the sampled subgraph G_s.
- $P(G)$: Distribution of the original graph G.

The goal is to show:

$$D_{\text{JS}}(P(G_s^{\text{with features}}), P(G)) \leq D_{\text{JS}}(P(G_s^{\text{without features}}), P(G)).$$

2. Sampling Subgraphs with and without Features

Subgraph Distribution:

- **Without node features:** $P(G_s^{\text{without features}})$ depends only on structural properties (e.g., edges, node degree). It neglects feature-based information, leading to a partial representation of the original graph.
- **With node features:** $P(G_s^{\text{with features}})$ includes both structural and feature-based information. This expanded representation better approximates $P(G)$, reducing divergence.

Key Observations:

- Neglecting node features creates a mismatch between $P(G_s)$ and $P(G)$ in the feature space, increasing divergence.
- Incorporating features ensures that the subgraph captures both structure and attributes of the original graph, making $P(G_s^{\text{with features}})$ closer to $P(G)$.

3. Proof Framework

(1) Subgraph Information Preservation

Using mutual information I as a measure of information preserved between G_s and G:

$$I(G_s^{\text{with features}}; G) \geq I(G_s^{\text{without features}}; G).$$

This inequality holds because node features provide additional information about G, increasing the overlap between $P(G_s)$ and $P(G)$.

(2) JS Divergence and Mutual Information

The JS divergence is inversely related to mutual information. Specifically, if I increases (i.e., more information is shared between distributions), the divergence decreases:

$$D_{\text{JS}}(P(G_s), P(G)) \propto -I(G_s; G).$$

Thus, the inclusion of node features increases $I(G_s; G)$, leading to:

$$D_{\text{JS}}(P(G_s^{\text{with features}}), P(G)) \leq D_{\text{JS}}(P(G_s^{\text{without features}}), P(G)).$$

(3) Reduced Feature-Space Divergence

When node features are excluded, the subgraph distribution $P(G_s^{\text{without features}})$ marginalises over features, effectively replacing X with a uniform or less informative distribution:

$$P(G_s^{\text{without features}}) = \int P(G_s|X)\, dX.$$

This marginalisation increases uncertainty and reduces the alignment between $P(G_s)$ and $P(G)$. By incorporating features:

$$P(G_s^{\text{with features}}) = P(G_s|X),$$

which retains the feature-space structure and reduces the divergence.

(4) KL Divergence Reduction

Consider the first KL term in the JS divergence:

$$D_{\text{KL}}(P(G_s) \| M),$$

where $M = \frac{1}{2}(P(G_s) + P(G))$. Including features reduces the difference between $P(G_s)$ and $P(G)$ in both the structural and feature spaces, thereby reducing $D_{\text{KL}}(P(G_s^{\text{with features}}) \| M)$ compared to $D_{\text{KL}}(P(G_s^{\text{without features}}) \| M)$.

Similarly, the second KL term $D_{\text{KL}}(P(G) \| M)$ also decreases due to the improved approximation of $P(G)$ by $P(G_s^{\text{with features}})$.