Beyond Localized Graph Neural Networks: An Attributed Motif Regularization Framework

Aravind Sankar*, Junting Wang*, Adit Krishnan, Hari Sundaram

University of Illinois at Urbana-Champaign, IL, USA

{asankar3, junting3, aditk2, hs1}@illinois.edu

Abstract—We present InfoMotif, a new semi-supervised, motif-regularized, learning framework over graphs. We overcome two key limitations of message passing in popular graph neural networks (GNNs): *localization* (a k-layer GNN cannot utilize features outside the k-hop neighborhood of the labeled training nodes) and *over-smoothed* (structurally indistinguishable) representations. We propose the concept of *attributed structural roles* of nodes based on their occurrence in different *network motifs*, independent of network proximity. Two nodes share attributed structural roles if they participate in topologically similar motif instances over co-varying sets of attributes. Further, InfoMotif achieves architecture independence by regularizing the node representations of arbitrary GNNs via *mutual information* maximization. Our training curriculum dynamically prioritizes multiple motifs in the learning process without relying on distributional assumptions in the underlying graph or the learning task. We integrate three state-of-the-art GNNs in our framework, to show significant gains (3–10% accuracy) across six diverse, real-world datasets. We see stronger gains for nodes with *sparse training labels* and *diverse attributes* in local neighborhood structures.

I. INTRODUCTION

This paper proposes a class of motif-regularized graph neural networks (GNNs); GNNs have emerged as a popular paradigm for semi-supervised learning on graphs due to their ability to learn representations combining topology and attributes. GNNs are typically formulated as a message passing framework [1], where the representation of a node is computed by a GNN layer aggregating features from its graph neighbors via learnable aggregators. Long-range dependencies are captured by using k layers to incorporate features from k-hop neighborhoods.

Localized message passing limitations: We illustrate two key limitations of prior k-layer GNN architectures: *k-hop localized* and *over-smoothed* representations (Figure 1).

1) GNNs, while highly expressive, are inherently *localized*: a k-layer GNN cannot utilize features of nodes that lie outside the k-hop neighborhood of the labeled training nodes. In Figure 1, nodes a and b belong to different classes. A 2-layer GNN sees unlabeled node c within the aggregation range of a (class 1) and outside the influence of b (class 2 and more than 2 hops away). Thus, a GNN will more likely label c with class 1 (than class 2). However, in reality, c and b display identical attributes (node color) in the local structure; a localized GNN fails to incorporate this factor.

2) GNNs with multiple layers learn *over-smoothed* node representations by iteratively aggregating neighbor features [2]. In Figure 1, nodes c and a share the same number of

neighbors with blue and green attributes; however, green neighbors of node a form triangles, while blue neighbors of node b (and c) form triangles. Considering *local nodal attribute arrangements*, node c is more similar to b than to a. The over-smoothing effect in GNNs obscures this attribute co-variation difference when classifying node c.

Thus, we require a new learning framework over graphs, to overcome the limitations of message passing in popular GNNs.
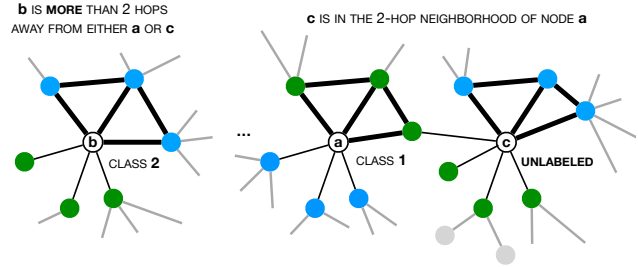

Fig. 1: **Localized message passing limitations:** A stylized example with a 2-layer GNN (colors indicate node attributes). Node a is in the 2-hop range of node c. Node c **does not receive** gradient updates from node b (class 2) since node b is more than 2 hops away. The GNN will likely label node c as class 1. Notice that c is in class 2 as since c and b have identical local structure and attribute co-variation.

One way to overcome these limitations is the paradigm of role discovery [3] that identifies nodes with structurally similar neighborhoods. In contrast to the notion of communities defined by network proximity, structural roles characterize nodes by their local connectivity and subgraph patterns independent of their location in the network [4]; thus, two nodes with similar roles may lie in different parts of the graph. Prior role-aware models learn similar representations for structurally similar nodes while ignoring nodal attributes [5], *i.e.*, they will assign the same role to nodes a and b in Figure 1 with topologically identical local structures; however, nodes a and b differ in their local attribute arrangements (blue vs. green attributes in triangles), and thus belong to different classes.

Present Work: To enable the expressivity to distinguish attributed structures, we propose the concept of *attributed structural roles* that identify structurally similar nodes with co-varying attributes, independent of network proximity. We ground structural roles on *network motifs*[1], which are higher-

---

[1]The terms network motif, graphlet, and induced subgraph are used interchangeably in literature.

*Equal contribution

order structures crucial to the organization of complex networks [6]. We define two nodes as sharing attributed structural roles if they participate in topologically similar motif instances over co-varying sets of attributes. We note that attribute co-variance permits for multiple discrete and continuous attributes, rather than stricter notions such as regular equivalence [4].

We propose InfoMotif, a GNN architecture-agnostic regularization framework that exploits the co-variance of attributes and motif structures. InfoMotif learns regularizers based on a set of network motifs, which vary in their task-specific significance. Specifically, across instances of the same motif (*e.g.*, a triangle structure), we learn discriminative attribute correlations to regularize the underlying GNN node representations; this encourages the GNN to learn statistical correspondences between distant nodes that participate in similarly attributed instances of that motif. We propose a novel training curriculum to integrate multiple motif regularizers while attending to motif types and skewed motif distributions. Our key contributions:

- **Attributed Structural Role Learning**: We propose the novel concept of *attributed structural roles* to regularize GNN models for semi-supervised learning. In contrast to prior work that identify structurally similar nodes agnostic to attributes [5], we use the principle of *mutual information maximization* to regularize node representations to capture *attribute correlations* in motif structures. InfoMotif unifies the expressive local neighborhood aggregation power of GNNs with the paradigm of structural role discovery.

- **Architecture-agnostic Regularization Framework:** To the best of our knowledge, InfoMotif is the first to address the limitations of localized message passing in GNNs through an architecture-agnostic framework. Unlike prior attempts that design new aggregators [7], [8], we achieve architecture independence by modulating the node representations learned by the base GNN, to capture attributed structural roles. We show significant gains over the state-of-the art GNNs.

- **Distribution-agnostic Multi-Motif Curriculum**: We propose two learning progress indicators, *task-driven utility* and *distributional novelty*, to integrate multiple motif regularizers within our framework. Unlike prior strategies [9], [10] that incorporate regularizers via tunable hyper-parameters, our training curriculum dynamically prioritizes different motifs in the learning process without relying on distributional assumptions on the underlying graph or on the learning task.

We integrate three state-of-the-art GNN models in our framework, to show significant gains (3-10% accuracy) with motif-based regularization on two diverse classes of datasets: *citation* networks that exhibit strong homophily and *air-traffic* networks that depend on structural roles. Our qualitative analysis indicates stronger gains for nodes with *sparse training labels* and *diverse attributes* in local neighborhood structures.

We organize the rest of the paper as follows. In Section II, we present the problem formulation, and introduce preliminaries on GNNs and network motifs. We describe our proposed framework InfoMotif in Sections III and IV, present experimental results in Section V, finally concluding in Section VII.

II. PRELIMINARIES

In this section, we formalize semi-supervised node classification on graphs via Graph Neural Networks and introduce network motifs to regularize the classification.

A. *Problem Definition*

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an attributed graph, with nodes $\mathcal{V}$ and edges $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$. Note, $\mathcal{V} = \mathcal{V}_L \cup \mathcal{V}_U$, the sets of labeled ($\mathcal{V}_L$) and unlabeled ($\mathcal{V}_U$) nodes in the graph. Let $\mathcal{N}(v)$ denote the neighbor set of node $v \in \mathcal{V}$ in $\mathcal{G}$, and $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times F}$ denotes the attribute matrix with rows $\mathbf{x}_v \in \mathbb{R}^F$ for node $v \in \mathcal{V}$. Each labeled node $v \in \mathcal{V}_L$ belongs to one of C classes, encoded by a one-hot vector $\mathbf{y}_v \in \mathbb{B}^C$ ($\mathbb{B} = \{0, 1\}$). Our goal is to predict the labels of the unlabeled nodes $v \in \mathcal{V}_U$. This is the familiar transductive learning setup for node classification [9].

B. *Graph Neural Networks*

Graph Neural Networks (GNNs) use multiple layers to learn node representations. At each layer $l > 0$, where 0 is the input layer, GNNs compute a representation for node v by aggregating features from its neighborhood, through a learnable aggregator function $f_{\theta,l}$ per layer. Using k layers allows for the k-hop neighborhood of a node to influence its representation.

Let $\mathbf{h}_{v,l-1} \in \mathbb{R}^D$ denote the representation of node v in layer $l - 1$. The l-th layer follows a message passing rule:

$$\mathbf{h}_{v,l} = f_{\theta,l}\Big(\mathbf{h}_{v,l-1}, \{\mathbf{h}_{u,l-1}\}\Big), \quad u \in \mathcal{N}_v \tag{1}$$

Equation (1) says that the node embedding $\mathbf{h}_{v,l} \in \mathbb{R}^D$ for node v at the l-th layer is a non-linear aggregation $f_{\theta,l}$ of the embeddings from layer $l - 1$ of node v and the embeddings of immediate network neighbors $u \in \mathcal{N}(v)$ of node v. The function $f_{\theta,l}$ defines the message passing mechanism at layer l and we can use a variety of aggregator architectures, including graph convolution [11], graph attention [12], and pooling [13]. The node representation for v at the input layer is $\mathbf{h}_{v,0}$ (*i.e.*, $l = 0$), where $\mathbf{h}_{v,0} = \mathbf{x}_v$ and $\mathbf{x}_v \in \mathbb{R}^F$. We designate the representation of node v at the final GNN layer $\mathbf{h}_v \in \mathbb{R}^D$, as its **base GNN representation**. In this work, we use GNNs as a collective term for networks that operate over graphs using localized message passing, as opposed to spectral methods [14] that learn convolutional filters from the entire graph.

C. *Network Motifs*

Network motifs are a general class of higher-order connectivity patterns, with a history of use in network science [6], [15]. A motif has several topologically equivalent appearances in the network called *motif instances*. Prior work [16], [17] shows how to efficiently compute motif instances for large graphs.

Definition 1 (Network Motif): A network motif $M_t = (\mathcal{V}_t, \mathcal{E}_t)$ is a connected, induced subgraph consisting of a subset $\mathcal{V}_t \subset \mathcal{V}$ and $\mathcal{E}_t = \{e \in \mathcal{E} \mid e = (u, v), u, v \in \mathcal{V}_t\}$. Let k_t be the number of nodes in M_t; that is, $k_t = |\mathcal{V}_t|$. We assume that a graph has a set of unique associated motifs $\mathcal{M} = \{M_1, \ldots, M_T\}$.

Definition 2 (Motif Instance): Let I_t be an induced subgraph of $\mathcal{G}$. We define I_t to be a motif instance of M_t if I_t is

473

Symbol	Description
$\mathcal{M}$	Set $\{M_1, \ldots, M_T\}$ of T network motifs
$\mathcal{I}_v(M_t)$	Set of instances of motif M_t in $\mathcal{G}$ that contain node v
$\mathbf{h}_{v,l}$	Representation of node v at layer l of GNN
$\mathbf{h}_v$	Base GNN representation of node v (final layer)
$\mathbf{h}_v^t$	Motif-gated representation of node v for motif M_t
$\mathbf{e}_{v,I_t}$	Instance-specific representation of v in $I_t \in \mathcal{I}_v(M_t)$
$\mathbf{s}_{v,t}$	Motif-level representation of node v for motif M_t
$\mathbf{z}_v$	Final Representation of node v
α_{vt}	Task-specific importance of motif M_t to node v
β_v	Novelty score for training node $v \in \mathcal{V}_L$

TABLE I: Notation

isomorphic to M_t. A motif M_t can have several motif instances in $\mathcal{G}$. While each such motif instance has a unique node set, two motif instances can share nodes. We denote the set of unique instances of M_t in $\mathcal{G}$ that contain node v as $\mathcal{I}_v(M_t)$.

In this work, we consider 3-node connected network motifs, *e.g.*, Figure 2 shows all 3-node, topologically distinct, directed (*e.g.*, citations) and undirected, connected network motifs.
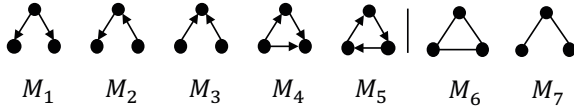

Fig. 2: Topologically distinct, directed (M_1 to M_5) and undirected (M_6 to M_7) 3-node, connected, network motifs

D. *Regularization*

We plan to use these local structural properties (*i.e.*, network motifs) to regularize the graph neural model during training. Typically, we train GNNs by minimizing the cross-entropy loss L_B, between model predictions $\hat{\mathbf{y}}_v \in \mathbb{R}^C$ and ground-truth labels $\mathbf{y}_v \in \mathbb{B}^C$ of training nodes in $v \in \mathcal{V}_L$, defined by:

$$L_B = - \sum_{v \in \mathcal{V}_L} \sum_{c=1}^{C} y_{v,c} \log \hat{y}_{v,c} \qquad (2)$$

where, the c-th index of the one-hot vector $\hat{y}_{v,c}$ refers to the probability that v belongs to the true class c. Notice that the loss L_B is agnostic to any local structural properties (*e.g.*, mixing patterns in social networks [18]) that may be indicative of the true node class. Thus, we develop a modified loss $L_B' = L_B + \lambda L_R$, where L_R is the regularization loss that incorporates attributed motif structure and λ is a constant. Our goal is to design L_R to overcome the two limitations of message-passing models: localized and over-smoothed node representations.

III. INFOMOTIF FRAMEWORK

In this section, we first discuss the structural properties of GNNs to motivate the notion of attributed structural roles. In section III-B, we present our motif-based mutual information maximization framework InfoMotif to regularize GNNs based on a single motif. Finally, in section III-C, we introduce our overall framework with a novel multi-motif training curriculum.

A. *Attributed Structural Role Learning*

A k-layer GNN computes a localized representation $\mathbf{h}_{v,k}$ for each node v that incorporates information from its k-hop neighborhood, denoted by $\mathcal{N}_k(v)$. For a node set $S \subseteq \mathcal{V}$,

let $\mathcal{N}_k(S) = \bigcup_{v \in S} \mathcal{N}_k(v)$ define its k-hop neighborhood, and $\mathbf{X}(S)$ denote its set of input node features. Let $\mathbf{Y}(\mathcal{V}_L)$ comprise the training labels of nodes in the labeled set $\mathcal{V}_L$. For a k-layer GNN trained on $\mathcal{V}_L$ using loss L_B (Equation (2)), let $\Theta^* = \{\Theta_1, \ldots, \Theta_k\}$ be the optimal parameters computed by its training algorithm. Now, we have the following proposition.

Proposition 3.1: Θ^* is a function of $\mathbf{X}(\mathcal{N}_k(\mathcal{V}_L))$, $\mathbf{Y}(\mathcal{V}_L)$ and changes in inputs $\mathbf{X}(\mathcal{V} \setminus \mathcal{N}_k(\mathcal{V}_L))$ will not affect Θ^*.

Proof Sketch. By an induction argument, the loss L_B can be written as $g(\Theta_1, \ldots, \Theta_k, \mathbf{Y}(\mathcal{V}_L), \mathbf{X}(\mathcal{N}_k(\mathcal{V}_L)))$ for some function $g(\cdot)$. Thus, when the GNN is trained on L_B using gradient updates, Θ^* must be independent of $\mathbf{X}(\mathcal{V} \setminus \mathcal{N}_k(\mathcal{V}_L))$.

Note that addition of a standard regularization term (*e.g.*, L_1 or L_2) only impacts $\{\Theta_1, \ldots, \Theta_k\}$; the overall loss still remains independent of $\mathcal{V} \setminus \mathcal{N}_k(\mathcal{V}_L)$, satisfying proposition 3.1.

Thus, the optimal parameters of a k-layer GNN are only affected by node features in the k-hop neighborhood $\mathcal{N}_k(\mathcal{V}_L)$ of the labeled set $\mathcal{V}_L$, *i.e.*, the features and connectivities of nodes in $\mathcal{V} \setminus \mathcal{N}_k(\mathcal{V}_L)$ are ignored in the training process.

Let the k-hop neighborhood of class c be $\mathcal{N}_k(\mathcal{V}_L(c))$ where $\mathcal{V}_L(c) = \{v \in \mathcal{V}_L : y_{vc} = 1\}$ is the set of nodes labeled with class c. Let $L_B(c)$ be the supervised loss term specific to class c. Now, the corollary directly follows from proposition 3.1:

Corollary 3.2: If node $v \notin \mathcal{N}_k(\mathcal{V}_L(c))$, the k-hop neighborhood of class c, then the loss $L_B(c)$ is independent of v.

The above corollary states that gradient updates from the supervised loss $L_B(c)$ for class c cannot reach nodes that lie outside the k-hop neighborhood of class c, *i.e.*, $\mathcal{N}_k(\mathcal{V}_L(c))$.

To illustrate its implications, we revisit Figure 1. Since node c lies beyond the 2-hop neighborhood of node b, node c does not affect the training loss at node b (which belongs to class 2). Thus, despite nodes c and b having identical co-variation of attributes and structure (blue neighbors form triangles), node c does not influence the training loss for all nodes with class 2.

B. *Single Motif Regularization*

In this section, we introduce InfoMotif, a framework to regularize node representations of the base GNN by exploiting the co-variance of node attributes and motif structures. We define *attributed structural roles* by assigning the same role to nodes that participate in motif instances over *co-varying* sets of attributes. In contrast, prior role-aware models [5] discover structurally similar nodes *agnostic* to attributes.

Now, we describe our regularization strategy to learn attribute co-variance for a single motif. In the next section, we extend these arguments to handle multiple motifs.

Motif-based Mutual Information: We first consider a single network motif type $M_t \in \mathcal{M}$ and a specific node $v \in \mathcal{V}$ to learn attribute co-variance across instances $\mathcal{I}_v(M_t)$ that contain v in the graph. To learn attributed structural roles, it is necessary to *contrast* the attributed instances of motif M_t against attributed node combinations that are not present in any instances of M_t.

We maximize the motif-based *mutual information* (MI) between a *motif-level* representation of v and corresponding *instance-specific* representations centered at v. By introducing motif-based MI maximization as a regularizer, the GNN is
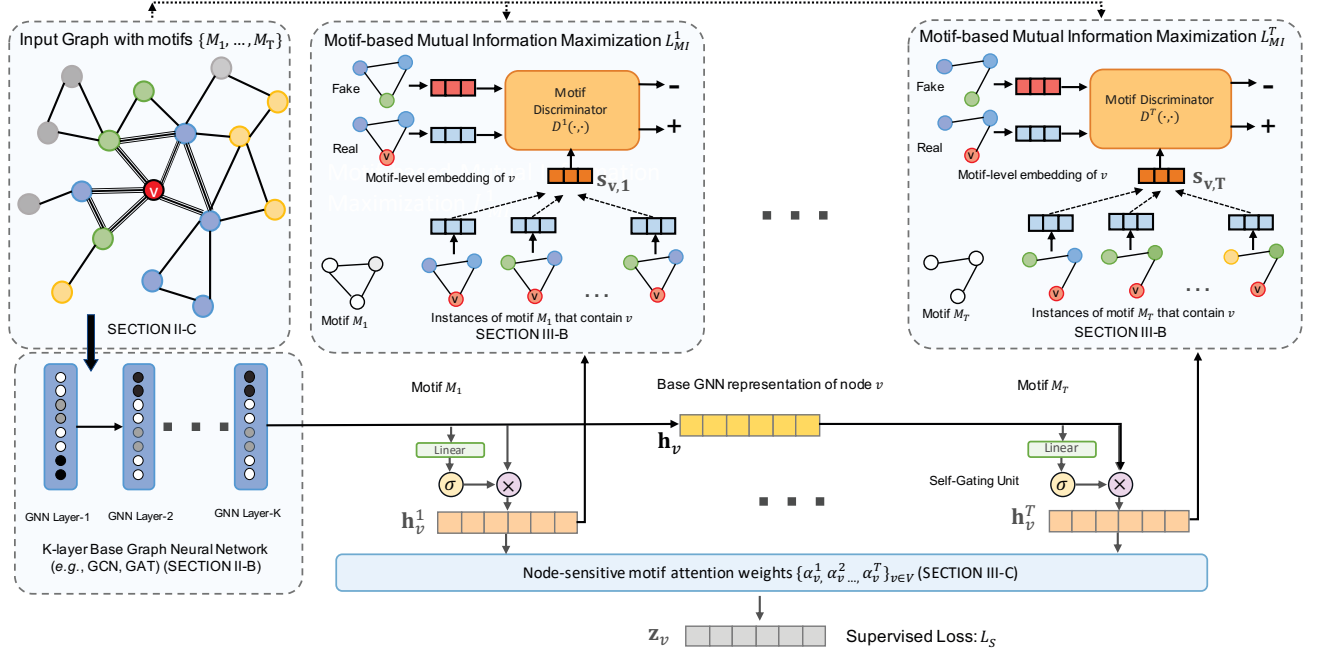
Fig. 3: Architecture diagram of InfoMotif depicting the model components: base GNN $f_{\theta,l}$ with k layers (bottom left), motif-based mutual information maximizing regularizers L_{MI}^t (top right), and attention module to compute final node representations $\mathbf{z}_v$ (bottom right). Instances of motif M_1 are shown in the graph (top left) with textured lines and colors indicate node attributes.

encouraged to learn *discriminative* statistical correspondences between nodes that participate in instances of the same motif.

We first adapt the base GNN representation $\mathbf{h}_v$ (see Section II-B), specific to motif M_t through a *motif gating* function $f_{\text{GATE}}^t : \mathbb{R}^D \mapsto \mathbb{R}^D$ resulting in a gated embedding $\mathbf{h}_v^t$. Then, we introduce a *motif instance encoder* $f_{\text{ENC}}^t : \mathbb{R}^D \times \mathbb{R}^{(k_t \times D)} :\mapsto \mathbb{R}^D$ to compute the instance-specific representation $\mathbf{e}_{v,I_t} \in \mathbb{R}^D$ of node v conditioned on other co-occurring nodes in instance $I_t \in \mathcal{I}_v(M_t)$. Finally, the motif-level representation $\mathbf{s}_{v,t} \in \mathbb{R}^D$ of node v summarizes the set of instance-specific representations $\{\mathbf{e}_{v,I_t}\}_{I_t \in \mathcal{I}_v(M_t)}$ through a permutation-invariant *motif readout* function $f_{\text{READ}}^t(\cdot)$, *e.g.*, averaging or pooling functions.

For each node $v \in \mathcal{V}$, we maximize motif-based mutual information L_{MI}^t between its instance-specific representations $\{\mathbf{e}_{v,I_t}\}_{I_t \in \mathcal{I}_v(M_t)}$ and motif-level representation $\mathbf{s}_{v,t}$, by defining I_{ψ^t} as a mutual information estimator for motif M_t that is *shared* across all nodes. The resulting objective is given by:

$$L_{MI}^t(\theta, \phi^t, \psi_t) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \sum_{I_t \in \mathcal{I}_v(M_t)} I_{\psi_t}(\mathbf{e}_{v,I_t}; \mathbf{s}_{v,t}) \quad (3)$$

where θ and ϕ^t denote the parameters of the layers $\{f_{\theta,l}\}_{l=1}^k$, and motif-specific transforms $\{f_{\text{GATE}}^t, f_{\text{ENC}}^t, f_{\text{READ}}^t\}$ respectively. By maximizing MI across all instances of motif M_t in the graph through a shared MI estimator I_ψ^t, we enable the GNN to learn correspondences between a pair of potentially distant nodes that participate in instances of motif M_t.

Mutual Information Maximization: Following neural MI estimation methods [19], [20], we model the estimator I_{ψ^t} as a *discriminator* network that learns a decision boundary to accurately distinguish between *positive* samples drawn from the joint distribution and *negative* samples drawn from the product of marginals. We train a *constrastive* discriminator network $\boldsymbol{D}_\psi^t : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}^+$, where $\boldsymbol{D}_\psi^t(\mathbf{e}_{v,I_t}, \mathbf{s}_{v,t})$ denotes the probability score assigned to this instance-motif pair. The positive samples $(\mathbf{e}_{v,I_t}, \mathbf{s}_{v,t})$ for $\boldsymbol{D}_\psi^t$ are the representations of observed instances $I_t \in \mathcal{I}_v(M_t)$ of motif M_t paired with the motif-level representation $\mathbf{s}_{v,t}$. The negative samples $(\mathbf{e}_{v,\widetilde{I}_t}, \mathbf{s}_{v,t})$ are derived by pairing $\mathbf{s}_{v,t}$ with the representations of negative instances $\widetilde{I}_t$ sampled from a distribution $P_{\mathcal{N}}(\widetilde{I}_t | M_t)$. The discriminator $\boldsymbol{D}_\psi^t$ is trained on a noise-contrastive objective L_{MI}^t between samples from the joint (positive pairs), and the product of marginals (negative pairs), which is defined as:

$$L_{MI}^t = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} L_{MI}^t(v) = -\frac{1}{2Q|\mathcal{V}|} \sum_{v \in \mathcal{V}} \sum_{i=1}^Q \Big[\mathbb{E}_{I_t} \log \boldsymbol{D}_\psi^t(\mathbf{e}_{v,I_t}, \mathbf{s}_{v,t})$$
$$+ \mathbb{E}_{\widetilde{I}_t} \log(1 - \boldsymbol{D}_\psi^t(\mathbf{e}_{v,\widetilde{I}_t}, \mathbf{s}_{v,t})) \Big] \quad (4)$$

where Q is the number of observed motif instances sampled per node. This objective maximizes MI between $\mathbf{s}_{v,t}$ and $\{\mathbf{e}_{v,I_t}\}_{I_t \in \mathcal{I}_v(M_t)}$ based on the Jensen-Shannon Divergence between their joint distribution and product of marginals [21].

We design the negative sampling distribution $P_{\mathcal{N}}(\widetilde{I}_t | M_t)$ to learn attribute co-variance in instances of motif M_t. For each positive instance I_t, the generated negative instance $\widetilde{I}_t$ is topologically equivalent but contains attributes that do not occur in instances of M_t in $\mathcal{G}$. By contrasting the observed instances of M_t against fake instances with perturbed attributes, $\boldsymbol{D}_\psi^t$ learns attributed structural roles with respect to motif M_t.

475

C. Multi-Motif Regularization Framework

Now, we extend our framework for any graph that includes a set of motifs $\mathcal{M} = \{M_1, \ldots, M_T\}$. A typical way to include regularizers (Equation (4)) from multiple motifs is given by:

$$L^{'} = L_B + \lambda L^{'}_{MI} = L_B + \lambda \cdot \frac{1}{T} \sum_{t=1}^{T} L_{MI}^t \qquad (5)$$

where λ is a tunable hyper-parameter to balance the supervised task loss L_B and motif regularizers. Intuitively, each motif $M_t \in \mathcal{M}$ is a connectivity pattern that can be viewed as defining one kind of structural role, *e.g.*, bridge nodes. Each motif has a different significance towards the learning task. Thus, a multi-motif framework should automatically identify the significance of different motifs without manual hand tuning.

In addition, real-world networks exhibit heavy-tailed degree and community distributions [22], which manifest as *skewed* (imbalanced) motif occurrences among nodes as well as across motif types. This further complicates the learning process of incorporating multiple motifs as regularizers. We identify three key aspects *task*, *node*, and *skew* for a multi-motif framework:

- **Task**: Distinguish the significance of different motifs to compute representations conditioned on the learning task.
- **Node**: Expressive power to control the extent of regularization exerted by each motif at a node-level granularity.
- **Skew**: Adapt to varying levels of motif occurrence skew without any distributional assumptions on the input graph.

To address these objectives, we first describe our approach to compute final node representations conditioned on multiple motifs, followed by two novel online reweighting strategies.

Task-driven Representations: The base GNN is trained by a supervised task loss L_B (Equation (2)) over the labeled node set $\mathcal{V}_L$. We instead aggregate the set of motif-gated representations ($\mathbf{h}_v^t$ for motif $M_t \in \mathcal{M}$), to compute the final representation $\mathbf{z}_v \in \mathbb{R}^D$ for node v. We learn attention weights α_{vt} to characterize the task-driven importance of motif M_t to node v and compute $\mathbf{z}_v$ through a weighted average, given by:

$$\mathbf{z}_v = \sum_{t=1}^{T} \alpha_{vt} \mathbf{h}_v^t \quad \alpha_{vt} = \frac{\exp\left(\boldsymbol{p} \cdot \mathbf{h}_v^t\right)}{\sum\limits_{t'=1}^{T} \exp\left(\boldsymbol{p} \cdot \mathbf{h}_v^{t'}\right)} \qquad (6)$$

where $\boldsymbol{p} \in \mathbb{R}^D$ defines the attention function and is learned by optimizing the final representations $\{\mathbf{z}_v\}_{v \in \mathcal{V}_L}$ of labeled nodes $\mathcal{V}_L$ using the supervised loss L_B (Equation (2)). The final representation $\mathbf{z}_v$ of each node $v \in \mathcal{V}$ is used for classification.

Node-sensitive Motif Regularization: Instead of using static uniform weights to incorporate motif regularizers (Equation (5)), we contextually weight the contributions of different motif regularization terms (Equation (4)) at a node-level granularity through the attention weights α_{vt} of motif M_t for node v.

$$L_{MI} = \frac{1}{nT} \sum_{t=1}^{T} \sum_{v \in \mathcal{V}} \alpha_{vt} L_{MI}^t(v) \qquad (7)$$

The loss L_{MI} varies the extent of regularization per node in proportion to the task-specific importance α_{vt} of motif M_t to node v. Notice that while the attention function is learned by

---

Algorithm 1 The framework of InfoMotif-GNN.

Input: Graph $\mathcal{G}$, Labeled node set $\mathcal{V}_L$, Base GNN $\{f_{\theta,l}\}_{l=1}^k$
Output: Motif-regularized embedding $\mathbf{z}_v$ for each node $v \in \mathcal{V}$

1: Initialize sample novelty weights $\beta_v = 1 \ \forall \ v \in \mathcal{V}_L$
2: **while** *not converged* **do**
3:    ▷ *Supervised loss over labeled node set $\mathcal{V}_L$*
4:    **for** each batch of nodes $\mathcal{V}_B \subseteq \mathcal{V}_L$ **do**
5:      Fix sample weights $\{\beta_v\}_{v \in \mathcal{V}_B}$ and optimize L_S on $\mathcal{V}_B$ using mini-batch gradient descent (Equation 9).
6:    **end for**
7:    Compute motif attention weights $\{\boldsymbol{\alpha_v}\}_{v \in \mathcal{V}}$ (Equation 6).
8:    ▷ *Motif-based InfoMax loss over entire node set $\mathcal{V}$*
9:    **for** each batch of nodes $\mathcal{V}_B \subseteq \mathcal{V}$ **do**
10:      Fix motif weights $\{\boldsymbol{\alpha}_v\}_{v \in \mathcal{V}}$ and optimize L_{MI} on $\mathcal{V}_B$ using mini-batch gradient descent (Equation 7)
11:    **end for**
12:    Compute sample weights $\{\beta_v\}_{v \in \mathcal{V}_L}$ (Equation 8).
13: **end while**
14: Compute $\mathbf{z}_v \in \mathbb{R}^D \ \forall \ v \in \mathcal{V}$ (Equation 6)

---

training the final representations $\mathbf{z}_v$ of labeled nodes $v \in \mathcal{V}_L$ on the supervised loss L_B, the motif-regularization loss L_{MI} (which operates on all nodes) re-weights each motif loss term per node with the estimated attention weights.

Skew-aware Sample Weighting: Prior work in curriculum and meta learning has shown the importance of re-weighting training examples to overcome training set biases [23]. In particular, re-weighting strategies that emphasize harder examples are effective at handling imbalanced data distributions [24]. We propose a *novelty-driven* re-weighting strategy to handle skew in motif occurrences across nodes and motif types.

The novelty β_v of node v is a function of its motif distribution, *i.e.*, novel nodes contain uncommon motif types in their neighborhood, which in turn reflects in their attention weight distribution over motifs. Let $\boldsymbol{\alpha}_v \in \mathbb{R}^T$ denote the vector of attention weights for a labeled node v over the motif set $\mathcal{M}$. Now, the novelty β_v of node v is high if its motif distribution $\boldsymbol{\alpha}_v$ significantly diverges from those of other nodes. We quantify β_v by the deviation (measured by euclidean distance) of $\boldsymbol{\alpha}_v$ from the mean motif distribution of labeled nodes $v \in \mathcal{V}_L$.

$$\beta_v = \frac{exp(\|\boldsymbol{\alpha}_v - \boldsymbol{\mu}\|^2)}{\sum\limits_{u \in \mathcal{V}_L} \exp(\|\boldsymbol{\alpha}_u - \boldsymbol{\mu}\|^2)} \quad \boldsymbol{\mu} = \frac{1}{|\mathcal{V}_L|} \sum_{v \in \mathcal{V}_L} \boldsymbol{\alpha}_v \quad (8)$$

The novelty scores are normalized over $\mathcal{V}_L$ using a softmax function, to give non-negative sample weights $0 < \beta_v \leq 1$. We now define the novelty-weighted supervised loss L_S as:

$$L_S = - \sum_{v \in \mathcal{V}_L} \beta_v \sum_{c=1}^{C} y_{vc} \log \hat{y}_{vc} \qquad (9)$$

In contrast to the original supervised loss L_B (Equation (2)), the re-weighted objective L_S induces a novelty-weighted training curriculum that progressively focuses on harder samples.

Model Training: The overall objective of InfoMotif is composed of two terms, the re-weighted supervised loss L_S

476

(Equation (9)), and motif regularizers (Equation (7)), given by:

$$L = L_S + \lambda L_{MI} \tag{10}$$

In practice, we optimize L_S and L_{MI} alternatively at each training epoch, which removes the need to tune balance hyperparameter λ. Algorithm 1 summarizes the training procedure.

Complexity Analysis: On the whole, the complexity of our model is $O(\boldsymbol{F}) + O(nTQD + nTD^2)$ where $O(\boldsymbol{F})$ is the base GNN complexity, T is the number of motifs, Q is sampled instance count per motif, and D the latent space dimensionality. Since $T \ll n$ and $Q \ll n$, the added complexity of our framework scales linearly with respect to the number of nodes.

IV. MODEL DETAILS

We now discuss the architectural details of our framework: motif instance encoder, gating, readout, and discriminator.

A. *Motif Gating*

We design a pre-filter with *self-gating units* (SGUs) to regulate information flow from the base GNN embedding $\mathbf{h}_v$ to the motif-based regularizer. The SGU $f_{\text{GATE}}^t(\cdot)$ for motif M_t learns a non-linear gate to modulate the input at a feature-wise granularity through dimension re-weighting, defined by:

$$\mathbf{h}_v^t = f_{\text{GATE}}^t(\mathbf{h}_v) = \mathbf{h}_v \odot \sigma(\mathbf{W}_g^t \mathbf{h}_v + \mathbf{b}_g^t) \tag{11}$$

where $\mathbf{W}^t \in \mathbb{R}^{D \times D}, \mathbf{b}^t \in \mathbb{R}^D$ are learned parameters, $\odot$ denotes the element-wise product operation, and σ is the sigmoid non-linearity. The self-gating mechanism effectively serves as a multiplicative skip-connection [25] that facilitates gradient flow from the motif-based regularizer to the GNN.

B. *Motif Instance Encoder*

The encoder $f_{\text{ENC}}(\cdot)$ computes the instance-specific representation $\mathbf{e}_{v,I_t}$ for node v conditioned on the gated representations $\{\mathbf{h}_u^t\}_{u \in I_t}$ of the nodes in instance I_t. We apply self-attentions [26] to compute a weighted average of the gated node representations $\{\mathbf{h}_u^t\}_{u \in I_t}$ in I_t. Specifically, f_{ENC} attends over each node $u \in I_t$ to compute attention weight α_u by comparing its gated representation $\mathbf{h}_u^t$ with that of node v, $\mathbf{h}_v^t$.

$$\mathbf{e}_{v,I_t} = \sum_{u \in I_t} \alpha_u \mathbf{h}_u^t \quad \alpha_u = \frac{\exp\left(\boldsymbol{a}^t \cdot [\mathbf{h}_u^t || \mathbf{h}_v^t]\right)}{\sum\limits_{u' \in I_t} \exp\left(\boldsymbol{a}^t \cdot [\mathbf{h}_{u'}^t || \mathbf{h}_v^t]\right)} \tag{12}$$

where $\boldsymbol{a}^t \in \mathbb{R}^{2D}$ is a weight vector parameterizing the attention function and $||$ denotes concatenation. We empirically find the self-attentional encoder to outperform other pooling alternatives.

C. *Motif Readout*

The readout function $f_{\text{READ}}^t(\cdot)$ summarizes the set of instance-specific representations $\{\mathbf{e}_{v,I_t}\}_{I_t \in \mathcal{I}_v(M_t)}$ to compute the motif-level representation $\mathbf{s}_{v,t}$. We use a simple averaging of instance-specific representations to define $f_{\text{READ}}^t(\cdot)$ as follows:

$$\mathbf{s}_{v,t} = f_{\text{READ}}^t\left(\{\mathbf{e}_{v,I_t}\}_{I_t \in \mathcal{I}_v(M_t)}\right) = \sigma\left(\sum_{I_t \in \mathcal{I}_v(M_t)} \frac{\mathbf{e}_{v,I_t}}{|\mathcal{I}_v(M_t)|}\right)$$

where σ is the sigmoid non-linearity. We adopt batch-wise training with motif instance sampling (~ 20 per node) to compute $\mathbf{s}_{v,t}$. Sophisticated readout architectures [27] are more likely necessary to handle larger sample sizes.

Dataset	Citation Networks			Air-Traffic Networks		
	Cora	Citeseer	Pubmed	Brazil	Europe	USA
# Nodes	2,485	2,110	19,717	131	399	1,190
# Edges	5,069	3,668	44,324	1,038	5,995	13,599
# Attributes	1,433	3,703	500	-	-	-
# Classes	7	6	3	4	4	4

TABLE II: Dataset statistics of three benchmark citation [29] and air-traffic [5] networks. Ground-truth classes in citation networks exhibit attribute homophily; ground-truth classes in flight networks indicate node structural roles.

D. *Motif Discriminator*

The discriminator D_ψ^t learns a motif-specific scoring function to assign higher likelihoods to observed instance-motif pairs relative to negative examples. Similar to prior work [21], [28], we use a bilinear scoring function defined by:

$$\boldsymbol{D}_\psi^t(\mathbf{e}_{v,I_t}, \mathbf{s}_v^t) = \sigma(\mathbf{e}_{v,I_t} \cdot \boldsymbol{W}_d^t \mathbf{s}_v^t) \tag{13}$$

where $\boldsymbol{W}_d^t \in \mathbb{R}^{D \times D}$ is a trainable scoring matrix and σ is the sigmoid non-linearity to convert raw scores into probabilities of $(\mathbf{e}_{v,I_t}, \mathbf{s}_v^t)$ being a positive example for motif M_t.

V. EXPERIMENTS

We present extensive quantitative and qualitative analyses on multiple diverse datasets. We first introduce datasets, baselines, and experimental setup (Section V-A, V-B, V-C, and V-D), followed by node classification results in Section V-D by integrating three GNN models in our framework. In Section V-F, we present a qualitative analysis to analyze the impact of *label sparsity and attribute diversity in local neighborhoods*. We then conduct an ablation study to understand our gains over the base GNN models in Section V-E, analyze parameter sensitivity in Section V-G and model efficiency in Section V-H, Finally, we discuss limitations and future directions in Section V-I.

A. *Datasets*

We conduct experiments on two diverse types of benchmark datasets: *citation networks* that exhibit strong homophily and *air-traffic networks* that depend on structural roles (Table II).

- **Citation Networks:** We consider three benchmark datasets, Cora, Citeseer, and PubMed [29], where nodes correspond to documents and edges represent citation links. Each document is associated with a bag-of-words feature vector and the task is to classify documents into different research topics.
- **Air-Traffic Networks:** We use three undirected networks Brazil, Europe, and USA [5] where nodes correspond to airports and edges indicate the existence of commercial flights. Class labels are assigned based on activity level, measured by the cardinality of flights or people that passed the airports. We use one-hot indicator vectors as node attributes. Notice that class labels are related to the role played by airports.

B. *Baselines*

We organize competing baselines into four categories based on whether they are *proximity-based vs. structural*; and the paradigm of *embedding learning vs. graph neural networks*:

Training Ratio	Data		Cora			Citeseer			PubMed		
	X	Y	20%	40%	60%	20%	40%	60%	20%	40%	60%
PROXIMITY-BASED GRAPH EMBEDDING METHODS											
Node2Vec [30]			75.7 ± 0.5	76.1 ± 0.5	77.6 ± 0.5	68.1 ± 0.5	69.1 ± 0.6	69.2 ± 0.4	80.1 ± 0.6	80.2 ± 0.6	80.4 ± 0.6
Motif2Vec [31]			79.0 ± 0.4	79.2 ± 0.4	79.8 ± 0.5	66.6 ± 0.4	67.1 ± 0.3	68.8 ± 0.5	79.8 ± 0.2	79.8 ± 0.4	79.9 ± 0.4
STRUCTURAL GRAPH EMBEDDING METHODS											
Struct2Vec [5]			35.4 ± 1.0	37.6 ± 1.3	39.0 ± 1.1	31.2 ± 0.8	35.1 ± 0.9	36.5 ± 0.7	48.5 ± 0.3	49.2 ± 0.4	49.6 ± 0.4
GraphWave [32]			39.5 ± 2.1	41.1 ± 1.5	42.2 ± 1.9	38.5 ± 1.2	40.6 ± 0.9	43.9 ± 1.0	43.0 ± 2.0	43.3 ± 1.3	44.3 ± 1.5
DRNE [33]			34.9 ± 1.5	36.5 ± 1.5	37.3 ± 1.6	30.8 ± 1.2	32.2 ± 1.2	34.6 ± 1.4	40.4 ± 0.7	41.6 ± 0.4	43.3 ± 0.5
STANDARD GRAPH NEURAL NETWORKS											
GCN [11]	✓	✓	81.6 ± 0.5	82.0 ± 0.4	83.0 ± 0.5	75.8 ± 0.5	76.6 ± 0.3	76.8 ± 0.4	85.7 ± 0.7	86.1 ± 0.5	86.4 ± 0.5
GAT [12]	✓	✓	80.9 ± 0.7	81.4 ± 0.2	81.8 ± 0.5	74.5 ± 0.7	75.5 ± 0.7	76.4 ± 0.5	83.3 ± 0.3	84.2 ± 0.3	84.3 ± 0.3
GraphSAGE [13]	✓	✓	81.3 ± 0.3	83.5 ± 0.3	84.2 ± 0.3	72.9 ± 0.3	73.8 ± 0.2	76.4 ± 0.4	86.6 ± 0.2	87.2 ± 0.3	88.0 ± 0.2
JKNet [8]	✓	✓	81.3 ± 0.8	83.6 ± 0.8	84.2 ± 0.8	71.5 ± 0.8	72.5 ± 0.7	73.3 ± 0.7	82.2 ± 0.4	83.8 ± 0.5	84.4 ± 0.4
DGI [21]	✓		76.2 ± 0.8	77.3 ± 0.9	78.2 ± 0.8	74.5 ± 0.7	74.7 ± 0.7	75.4 ± 0.7	78.2 ± 0.9	78.5 ± 0.9	79.5 ± 0.9
STRUCTURAL GRAPH NEURAL NETWORKS											
DemoNet [7]	✓	✓	81.0 ± 0.6	82.4 ± 0.5	83.4 ± 0.7	67.9 ± 0.7	68.5 ± 0.6	68.9 ± 0.6	79.5 ± 0.4	80.5 ± 0.4	81.3 ± 0.4
Motif-CNN [34]	✓	✓	81.6 ± 0.5	82.8 ± 0.5	83.2 ± 0.5	73.4 ± 0.3	76.8 ± 0.3	77.1 ± 0.3	87.3 ± 0.1	87.5 ± 0.1	88.2 ± 0.1
MCN [35]	✓	✓	81.1 ± 0.9	82.4 ± 0.8	83.1 ± 0.9	73.2 ± 0.4	75.9 ± 0.7	76.6 ± 0.6	85.2 ± 0.6	85.9 ± 0.5	86.4 ± 0.7
MOTIF-REGULARIZED GRAPH NEURAL NETWORKS (INFOMOTIF)											
InfoMotif-GCN	✓	✓	$\mathbf{85.7 \pm 0.4}$	$\mathbf{87.4 \pm 0.4}$	$\mathbf{88.2 \pm 0.3}$	$\mathbf{77.7 \pm 0.5}$	$\mathbf{78.5 \pm 0.5}$	$\mathbf{80.1 \pm 0.5}$	$\mathbf{87.5 \pm 0.2}$	$\mathbf{88.3 \pm 0.2}$	$\mathbf{88.7 \pm 0.4}$
InfoMotif-JKNet	✓	✓	85.5 ± 0.3	86.5 ± 0.5	88.0 ± 0.2	74.5 ± 0.8	76.7 ± 0.9	77.8 ± 0.9	87.0 ± 0.2	87.9 ± 0.3	88.2 ± 0.3
InfoMotif-GAT	✓	✓	85.5 ± 0.3	87.2 ± 0.7	88.0 ± 0.2	76.5 ± 0.5	77.0 ± 0.4	78.9 ± 0.4	85.9 ± 0.4	86.2 ± 0.5	86.3 ± 0.5

TABLE III: Node classification results (% test accuracy) on citation networks using 10 random train/validation/test splits per training ratio (20%, 40% and 60%). **X** and **Y** denote the use of node attributes and training labels respectively towards representation learning. We report mean accuracy and standard deviation over 5 trials. We show GraphSAGE results with the best performing aggregator. InfoMotif consistently improves results of all three base GNNs by 3.5% on average across datasets.

- **Proximity-based embedding methods**: Conventional methods, node2vec [30] that learns from second-order random walks, and motif2vec [31] that models higher-order proximity.
- **Structural embedding methods**: Structural role-aware models struc2vec [5], GraphWAVE [32], and DRNE [33].
- **Standard Graph Neural Networks**: State-of-the-art GNN models based on localized message passing: GCN [11], GraphSAGE [13], GAT [12], JK-Net [8] and DGI [21].
- **Structural Graph Neural Networks**: Motif-based Motif-CNN [34], MCN [35] and degree-specific DEMO-Net [7].

C. Experimental Setup

We tested InfoMotif by integrating GCN, JK-Net and GAT as base GNNs within our framework. We consider the largest connected component in each dataset, and use the set of all directed 3-node motifs in citation networks and undirected 3-node motifs in air-traffic networks (Figure 2). To fairly compare different models [36], we evaluate different train/validation/test splits (training ratios of 20%, 40%, and 60%). We create 10 random data splits per training ratio and report the mean test classification accuracy along with standard deviation.

All experiments were conducted on a Tesla K-80 GPU using PyTorch. Our implementation of InfoMotif is publicly available[2]. For citation networks, we use two-layer base GNNs with layer sizes of 256 each, while using 64 for the smaller air-traffic networks. We train the base JK-Net using 4 GCN

[2] https://github.com/CrowdDynamicsLab/InfoMotif

layers and maxpool layer aggregation, while the base GAT learns 8 attention heads per layer. The model is trained for a maximum of 100 epochs with a batch size of 256 nodes with Adam optimizer. We also apply dropout with a rate of 0.5, and tune the learning rate in the range $\{10^{-4}, 10^{-3}, 10^{-2}\}$.

D. Experimental Results

Our experimental results comparing InfoMotif with three base GNNs, against competing baselines on citation and air-traffic networks, are shown in Tables III and IV respectively.

In citation networks, GNNs generally outperform conventional methods. Moreover, attribute-agnostic structural embedding methods perform poorly and structural GNNs perform comparably to standard GNNs. Citation networks exhibit strong attribute homophily in local neighborhoods; thus, structural GNNs do not provide much benefits over state-of-the-art message-passing GNNs. In contrast, our framework InfoMotif regularizes GNNs to discover distant nodes with similar attributed structures across the entire graph. InfoMotif achieves consistent average accuracy gains of 3% for all three variants.

In air-traffic networks, structural embedding methods outperform their proximity-based counterparts, with a similar trend for structural GNNs. Here, class labels rely more on node structural roles than the labels of neighbors. JK-net outperforms competing GNNs, signifying the importance of long-range dependencies in air-traffic networks. InfoMotif enables GNNs to learn structural roles agnostic to network proximity, and achieves significant gains of 10% on average across all datasets.

Training Ratio	X	Y	USA 20%	USA 40%	USA 60%	Europe 20%	Europe 40%	Europe 60%	Brazil 20%	Brazil 40%	Brazil 60%
P R O X I M I T Y - B A S E D G R A P H E M B E D D I N G M E T H O D S											
Node2Vec [30]			24.6 ± 0.9	24.8 ± 0.9	25.6 ± 0.9	36.5 ± 1.0	37.4 ± 1.1	38.0 ± 1.0	26.3 ± 1.4	30.4 ± 1.3	33.9 ± 1.4
Motif2Vec [31]			51.3 ± 1.1	54.8 ± 1.1	55.0 ± 1.1	37.1 ± 1.2	38.1 ± 1.2	39.5 ± 1.1	27.2 ± 1.5	33.9 ± 1.5	35.7 ± 1.5
S T R U C T U R A L G R A P H E M B E D D I N G M E T H O D S											
Struct2Vec [5]			50.4 ± 0.8	51.3 ± 0.8	53.8 ± 0.8	42.5 ± 0.7	45.6 ± 0.8	48.8 ± 0.7	45.8 ± 1.1	51.8 ± 1.1	57.1 ± 1.1
GraphWave [32]			45.2 ± 1.4	48.0 ± 1.4	51.4 ± 1.5	38.1 ± 1.9	41.1 ± 1.6	42.1 ± 2.0	40.2 ± 2.0	43.1 ± 1.8	48.5 ± 2.2
DRNE [33]			51.3 ± 1.1	52.4 ± 1.1	53.3 ± 1.1	43.1 ± 1.7	47.6 ± 1.3	50.8 ± 1.6	46.5 ± 2.7	50.2 ± 2.3	58.1 ± 2.0
S T A N D A R D G R A P H N E U R A L N E T W O R K S											
GCN [11]	✓	✓	51.9 ± 0.9	56.0 ± 0.9	57.0 ± 0.8	37.4 ± 0.9	40.1 ± 0.8	41.0 ± 0.8	36.5 ± 1.5	38.9 ± 1.6	39.3 ± 1.4
GAT [12]	✓	✓	52.7 ± 1.0	53.5 ± 0.9	56.3 ± 0.9	31.5 ± 1.0	34.3 ± 1.0	38.0 ± 1.0	37.3 ± 1.6	37.9 ± 1.6	38.2 ± 1.7
GraphSAGE [13]	✓	✓	45.3 ± 1.2	49.4 ± 1.2	50.4 ± 1.1	28.8 ± 1.0	32.5 ± 1.0	37.9 ± 1.0	36.1 ± 1.6	37.5 ± 1.6	39.3 ± 1.7
JKNet [8]	✓	✓	53.8 ± 1.2	56.1 ± 1.0	61.3 ± 1.0	49.7 ± 1.1	53.8 ± 1.1	54.3 ± 1.2	55.9 ± 1.5	58.4 ± 1.8	60.0 ± 1.4
DGI [21]	✓		46.4 ± 1.3	47.3 ± 1.2	48.1 ± 1.2	37.5 ± 1.5	39.9 ± 1.5	42.3 ± 1.4	41.4 ± 1.6	45.2 ± 1.7	44.1 ± 1.5
S T R U C T U R A L G R A P H N E U R A L N E T W O R K S											
DemoNet [7]	✓	✓	58.6 ± 1.2	58.8 ± 1.1	61.3 ± 1.0	40.4 ± 1.3	46.2 ± 1.2	47.5 ± 1.2	46.1 ± 1.4	48.9 ± 1.5	49.2 ± 1.5
Motif-CNN [34]	✓	✓	53.6 ± 1.0	54.2 ± 1.0	55.6 ± 0.9	37.9 ± 1.0	41.1 ± 1.1	42.8 ± 1.0	28.9 ± 1.6	35.7 ± 1.7	39.3 ± 1.7
MCN [35]	✓	✓	54.8 ± 1.4	54.9 ± 1.3	55.3 ± 1.1	36.8 ± 1.2	39.6 ± 1.5	41.2 ± 1.4	42.9 ± 1.6	43.6 ± 1.4	47.2 ± 1.5
M O T I F - R E G U L A R I Z E D G R A P H N E U R A L N E T W O R K S (I N F O M O T I F)											
InfoMotif-GCN	✓	✓	59.5 ± 0.9	62.9 ± 0.7	65.0 ± 0.7	$\mathbf{53.5 \pm 0.6}$	56.9 ± 0.6	$\mathbf{58.8 \pm 0.7}$	56.6 ± 1.2	60.7 ± 1.2	67.9 ± 1.1
InfoMotif-JKNet	✓	✓	$\mathbf{61.8 \pm 1.6}$	$\mathbf{64.3 \pm 1.2}$	$\mathbf{67.5 \pm 1.5}$	53.1 ± 1.2	$\mathbf{56.9 \pm 0.6}$	57.5 ± 1.2	$\mathbf{62.7 \pm 1.8}$	$\mathbf{67.9 \pm 1.5}$	$\mathbf{80.4 \pm 1.9}$
InfoMotif-GAT	✓	✓	58.0 ± 0.4	60.4 ± 0.3	62.6 ± 0.7	46.0 ± 1.5	50.0 ± 2.0	56.3 ± 0.5	50.6 ± 1.3	56.3 ± 1.1	58.9 ± 1.3

TABLE IV: Node classification results (% test accuracy) on air-traffic networks. Structural embedding methods and GNNs typically outperform proximity-based models. InfoMotif JK-Net achieves significant gains of 4% to 14% across datasets.

E. Ablation Study

We present an ablation study on citation networks to analyze the importance of major components in InfoMotif (Table V)

- **Remove novelty-driven sample weighting**. We set the novelty $\beta_v = 1$ (Equation (9)) to test the importance of addressing motif occurrence skew. We observe consistent 1% gains due to our novelty-driven sample weighting.
- **Remove task-driven motif weighting**. We remove the node-sensitive motif weights from the motif regularization loss (Equation (7)) by setting $\alpha_{vt} = 1$ for every node-motif pair. Contextually weighting different motif regularizers at a node-level granularity results in 2% average accuracy gains.
- **Remove both novelty and task driven weighting**. This variant applies a uniform motif regularization over all nodes without distinguishing the nodes-sensitive relevance of each motif, which significantly degrades classification accuracy.

Dataset	Cora	Citeseer	Pubmed
InfoMotif-GCN ($L_S + \lambda L_{MI}$)	$\mathbf{87.4 \pm 0.4}$	$\mathbf{78.5 \pm 0.5}$	$\mathbf{88.3 \pm 0.2}$
w/o novelty weights ($\beta_v = 1$ in eq. (9))	86.4 ± 0.5	77.6 ± 0.5	87.8 ± 0.3
w/o task weights ($\alpha_{vt} = 1$ in eq. (7))	84.6 ± 0.4	77.3 ± 0.4	87.3 ± 0.2
w/o novelty and task weights	84.0 ± 0.5	76.4 ± 0.6	87.3 ± 0.2
Base model GCN (L_B)	82.0 ± 0.4	76.6 ± 0.3	86.1 ± 0.5

TABLE V: Ablation study results with 40% training ratio on citation networks. The novelty and task weighting strategies improve classification accuracies by 2% on average.

F. Qualitative Analysis

We qualitatively examine the source of InfoMotif's gains over the base GNN (GCN due to its consistent performance).

by analyzing *label sparsity* and *attribute diversity* in local node neighborhoods, on the Cora and Citeseer citation networks.

Label Sparsity: We define the *label fraction* for a node as the fraction of labeled training nodes in its 2-hop neighborhood, *i.e.*, a node exhibits label sparsity if it has very few or no labeled training nodes within its 2-hop aggregation range. We separate test nodes into four quartiles by their label fraction. Figure 4 depicts classification results for GCN and InfoMotif-GCN under each quartile (Q1 has nodes with small label fractions).

InfoMotif has stronger performance gains over GCN for nodes with smaller label fractions (quartiles Q1 and Q2), which empirically validates the efficacy of our motif-based regularization framework in addressing the key limitation of GNNs (Section III-A), *i.e.*, InfoMotif benefits nodes with very few or no labeled nodes within their k-hop aggregation ranges.

Attribute Diversity: We measure the local *attribute diversity* of a node by the mean pair-wise attribute dissimilarity (computed by cosine distance) of itself with other nodes in its 2-hop neighborhood, *i.e.*, a node that exhibits strong homophily with its neighbors has low attribute diversity. We report classification results across attribute diversity quartiles in Figure 5.

Nodes with diverse attributed neighborhoods are typically harder examples for classification. Regularizing GNNs to learn attributed structures via motif occurrences can accurately classify diverse nodes, as evidenced by the higher relative gains of InfoMotif for diverse nodes (quartiles Q3 and Q4).

G. Parameter Sensitivity

We examine the effect of hyper-parameter Q that controls the number of motif instances sampled per node to train our motif-
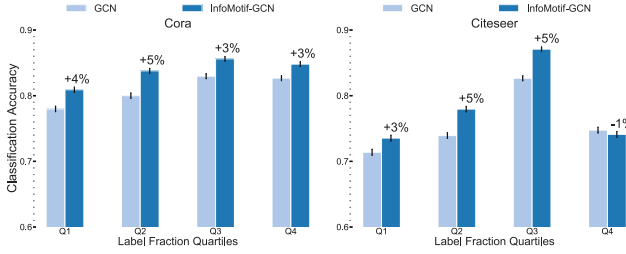
479

Fig. 4: Classification accuracy over *label fraction* quartiles. (Q1: smaller label fraction). InfoMotif has larger gains over GCN in Q1 & Q2 (nodes that exhibit label sparsity)
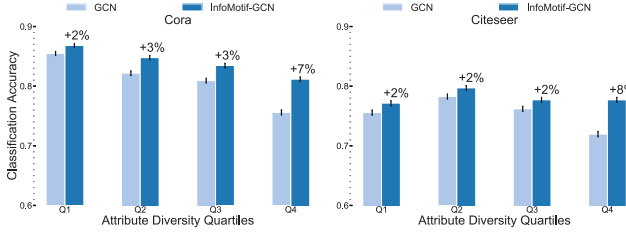

Fig. 5: Classification accuracy across attribute diversity quartiles. (Q4: high attribute diversity). InfoMotif has stronger gains in Q3 & Q4 (nodes with diverse attributed neighborhoods).

based discriminators (Equation (4)). Figure 6 shows variation in accuracies of our three GNN variants with the number of sampled instances (5 to 30), on Cora and Citeseer networks.

Performance of all GNN variants stabilize with 20 instances across both datasets. Since the complexity of our framework scales linearly with Q, we fix $Q = 20$ across datasets to provide an effective trade-off between compute-cost and performance
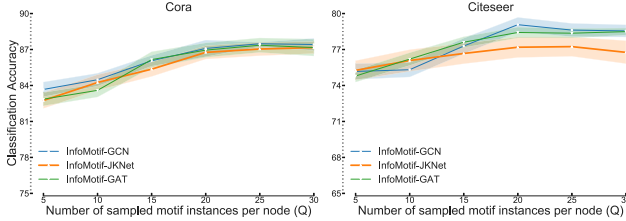
Fig. 6: Classification accuracy increases slowly with the number of sampled motif instances and stabilizes around 15 to 20. Variance bands indicate 95% confidence intervals over 10 runs.

H. *Efficiency Analysis*

We empirically evaluate the added complexity of InfoMotif on two GNN models, GCN and GAT. We report the time per epoch on synthetically generated Barabasi-Albert networks [37] with 5000 nodes and increasing link density (Figure 7).

InfoMotif adds a small fraction of the base GNN runtime, and the added complexity scales linearly with the number of nodes, as evidenced by its nearly constant runtime gap over increasing link density (Figure 7). Furthermore, our GCN variant InfoMotif-GCN is significantly more efficient than GAT.

I. *Discussion*

Our framework is orthogonal to advances in GNN architectures that enhance the structural distinguishability of node
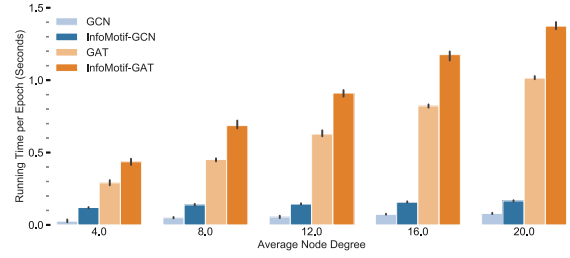

Fig. 7: Runtime comparison of InfoMotif variants with its base GNNs. InfoMotif has minimal computational overheads; notice the nearly constant runtime gap with increasing node degree.

representations through carefully designed aggregators. We regularize arbitrary GNNs to learn statistical correspondences between distant nodes with co-varying attribute structures. Our abstraction of roles through motifs generalizes across diverse types of networks, *e.g.*, signed and heterogeneous motifs [38].

Our key hypothesis is the importance of attribute co-variance in local structures towards the learning application (*e.g.*, classification in social networks). Our substantial gains on two diverse classes of datasets indicates broad applicability for InfoMotif across networks with varied structural characteristics. However, the gains may diminish in application scenarios where learning such co-variance is not beneficial or even necessary.

VI. RELATED WORK

GNNs learn node representations by recursively aggregating features from local neighborhoods in an end-to-end manner, with diverse applications, including information diffusion prediction [39], social recommendation [40], and community question answering [41]. Graph Convolutional Networks (GCNs) [11] learn degree-weighted aggregators, which can be interpreted as a special form of Laplacian smoothing [2]. Many models generalize GCN with a wide range of aggregators, *e.g.*, self-attentions [12], [42], mean and max pooling functions [13], etc. However, all these models learn node representations that inherently overfit to the k-hop neighborhood around each node.

There are two broad categories of techniques that capture contributions from distant nodes for graph representation learning: *non-local GNNs*, and *structural role-based embeddings*.

Non-local methods expand the propagation range of GNNs to aggregate node representations of differing localities, *e.g.*, JKNet [8] uses skip-connections to vary the influence radius per node, PGNN [43] captures global network positions via shortest-paths, and DGI [21] maximizes MI between node representations and a summary representation of the entire graph. However, they either operate on a local scale [44], or learn coarse structural properties, which limits their ability to capture features from distant yet structurally similar nodes.

Role-aware models embed structurally similar nodes close in the latent space, independent of network position [4], [45]. A few approaches [33] employ strict definitions of structural equivalence to embed nodes with identical local structures to the same point in the latent space, while others utilize structural node features (*e.g.*, node degrees, motif count statistics) to extend classical proximity-preserving embedding methods,

e.g., feature-based matrix factorization [46] and random walk methods [5]. Notably, a few methods design structural GCNs via motif adjacency matrices [34], [35], [47]. However, all these methods model structural roles without considering node attributes. InfoMotif is different since we regularize GNNs based on the co-variance of attributes and motif structures.

A related direction is higher-order network representation learning that models proximity via network motifs [31]. However, such representations are still highly localized and cannot identify structurally similar nodes independent of network proximity. In contrast, we contrastively learn attribute correlations in motifs to identify correspondences between distant nodes.

VII. CONCLUSION

This paper presents a new class of motif-regularized GNNs with an architecture-agnostic framework InfoMotif for semi-supervised learning on graphs. To overcome limitations of prior GNNs due to localized message passing, we introduce attributed structural roles to regularize GNNs by learning statistical dependencies between structurally similar nodes with co-varying attributes, independent of network proximity. InfoMotif maximizes motif-based mutual information, and dynamically prioritizes the significance of different motifs. Our experiments on six real-world datasets show substantial consistent gains for InfoMotif over state-of-the-art methods.

REFERENCES

[1] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE TNNLS*, 2020.

[2] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *AAAI*, 2018.

[3] R. A. Rossi and N. K. Ahmed, "Role discovery in networks," *TKDE*, vol. 27, no. 4, pp. 1112–1131, 2014.

[4] R. A. Rossi, D. Jin, S. Kim, N. K. Ahmed, D. Koutra, and J. B. Lee, "From community to role-based graph embeddings," *arXiv preprint arXiv:1908.08572*, 2019.

[5] L. F. Ribeiro, P. H. Saverese, and D. R. Figueiredo, "struc2vec: Learning node representations from structural identity," in *KDD*, 2017, pp. 385–394.

[6] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[7] J. Wu, J. He, and J. Xu, "Demo-net: Degree-specific graph neural networks for node and graph classification," in *KDD*. ACM, 2019, pp. 406–415.

[8] K. Xu, C. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," in *ICML*, 2018, pp. 5453–5462.

[9] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *NIPS*, 2004, pp. 321–328.

[10] C. Zhuang and Q. Ma, "Dual graph convolutional networks for graph-based semi-supervised classification," in *WWW*, 2018, pp. 499–508.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[12] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *ICLR*, 2018.

[13] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NIPS*, 2017, pp. 1024–1034.

[14] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *ICLR*, 2014.

[15] A. Paranjape, A. R. Benson, and J. Leskovec, "Motifs in temporal networks," in *WSDM*. ACM, 2017, pp. 601–610.

[16] R. A. Rossi, R. Zhou, and N. K. Ahmed, "Estimation of graphlet counts in massive networks," *IEEE TNNLS*, vol. 30, no. 1, pp. 44–57, 2018.

[17] P. Ribeiro, P. Paredes, M. E. Silva, D. Aparicio, and F. Silva, "A survey on subgraph counting: Concepts, algorithms and applications to network motifs and graphlets," *arXiv preprint arXiv:1910.13011*, 2019.

[18] L. Peel, J.-C. Delvenne, and R. Lambiotte, "Multiscale mixing patterns in networks," *PNAS*, vol. 115, no. 16, pp. 4057–4062, 2018.

[19] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, D. Hjelm, and A. Courville, "Mutual information neural estimation," in *ICML*, 2018, pp. 530–539.

[20] D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *ICLR*, 2019.

[21] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep Graph Infomax," in *ICLR*, 2019.

[22] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[23] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *ICML*, 2018.

[24] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019, pp. 9268–9277.

[25] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *ICML*, 2017, pp. 933–941.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[27] O. Vinyals, S. Bengio, and M. Kudlur, "Order matters: Sequence to sequence for sets," *arXiv preprint arXiv:1511.06391*, 2015.

[28] A. Sankar, Y. Wu, Y. Wu, W. Zhang, H. Yang, and H. Sundaram, "Groupim: A mutual information maximization framework for neural group recommendation," in *SIGIR*, 2020, pp. 1279–1288.

[29] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI magazine*, vol. 29, no. 3, pp. 93–93, 2008.

[30] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *KDD*, 2016, pp. 855–864.

[31] M. R. Dareddy, M. Das, and H. Yang, "motif2vec: Motif aware node representation learning for heterogeneous networks," in *BigData*. IEEE, 2019, pp. 1052–1059.

[32] C. Donnat, M. Zitnik, D. Hallac, and J. Leskovec, "Learning structural node embeddings via diffusion wavelets," in *KDD*, 2018, pp. 1320–1329.

[33] K. Tu, P. Cui, X. Wang, P. S. Yu, and W. Zhu, "Deep recursive network embedding with regular equivalence," in *KDD*, 2018, pp. 2357–2366.

[34] A. Sankar, X. Zhang, and K. C.-C. Chang, "Motif-based convolutional neural network on graphs," *arXiv preprint arXiv:1711.05697*, 2017.

[35] J. B. Lee, R. A. Rossi, X. Kong, S. Kim, E. Koh, and A. Rao, "Graph convolutional networks with motif-based attention," in *CIKM*, 2019, pp. 499–508.

[36] O. Shchur, M. Mumme, A. Bojchevski, and S. Günnemann, "Pitfalls of graph neural network evaluation," *NeurIPS Workshop*, 2018.

[37] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[38] R. A. Rossi, N. K. Ahmed, A. Carranza, D. Arbour, A. Rao, S. Kim, and E. Koh, "Heterogeneous network motifs," *arXiv preprint arXiv:1901.10026*, 2019.

[39] A. Sankar, X. Zhang, A. Krishnan, and J. Han, "Inf-vae: A variational autoencoder framework to integrate homophily and influence in diffusion prediction," in *WSDM*, 2020, pp. 510–518.

[40] A. Krishnan, H. Cheruvu, C. Tao, and H. Sundaram, "A modular adversarial approach to social recommendation," in *CIKM*. ACM, 2019, pp. 1753–1762.

[41] K. Narang, C. Yang, A. Krishnan, J. Wang, H. Sundaram, and C. Sutter, "An induced multi-relational framework for answer selection in community question answer platforms," *arXiv preprint arXiv:1911.06957*, 2019.

[42] A. Sankar, Y. Wu, L. Gou, W. Zhang, and H. Yang, "Dysat: Deep neural representation learning on dynamic graphs via self-attention networks," in *WSDM*, 2020, pp. 519–527.

[43] J. You, R. Ying, and J. Leskovec, "Position-aware graph neural networks," in *ICML*, 2019, pp. 7134–7143.

[44] Z. Peng, W. Huang, M. Luo, Q. Zheng, Y. Rong, T. Xu, and J. Huang, "Graph representation learning via graphical mutual information maximization," in *WWW*, 2020, pp. 259–270.

[45] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li, "Rolx: structural role extraction & mining in large graphs," in *KDD*, 2012, pp. 1231–1239.

[46] R. A. Rossi, N. K. Ahmed, E. Koh, S. Kim, A. Rao, and Y. Abbasi-Yadkori, "A structural graph representation learning framework," 2020.

[47] A. Sankar, X. Zhang, and K. C.-C. Chang, "Meta-gnn: Metagraph neural network for semi-supervised learning in attributed heterogeneous information networks," *ASONAM. IEEE*, pp. 137–144, 2019.