Knowledge and Information Systems (2022) 64:2091–2121
https://doi.org/10.1007/s10115-022-01694-5

REGULAR PAPER

Self-supervised role learning for graph neural networks

Aravind Sankar[1] · Junting Wang[1] · Adit Krishnan[1] · Hari Sundaram[1]

Received: 4 February 2021 / Revised: 15 May 2022 / Accepted: 21 May 2022 /
Published online: 13 July 2022

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 formulate *attributed structural roles* of nodes based on their occurrence in different *network motifs*, independent of network proximity. Network motifs are higher-order structures indicating connectivity patterns between nodes and are crucial to the organization of complex networks. Two nodes share attributed structural roles if they participate in topologically similar motif instances over covarying sets of attributes. InfoMotif achieves architecture-agnostic regularization of arbitrary GNNs through novel self-supervised learning objectives based on *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 notable performance gains (3–10% accuracy) across nine diverse real-world datasets spanning homogeneous and heterogeneous networks. Notably, we see stronger gains for nodes with *sparse training labels* and *diverse attributes* in local neighborhood structures.

Keywords Association rule · Data mining · Itemset · Transaction collection

1 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, without relying on expensive feature engineering. GNNs are typically formulated as a message passing framework [62], where the representation of a node is computed by a GNN layer aggregating features from

✉ Aravind Sankar
asankar3@illinois.edu

[1] Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA

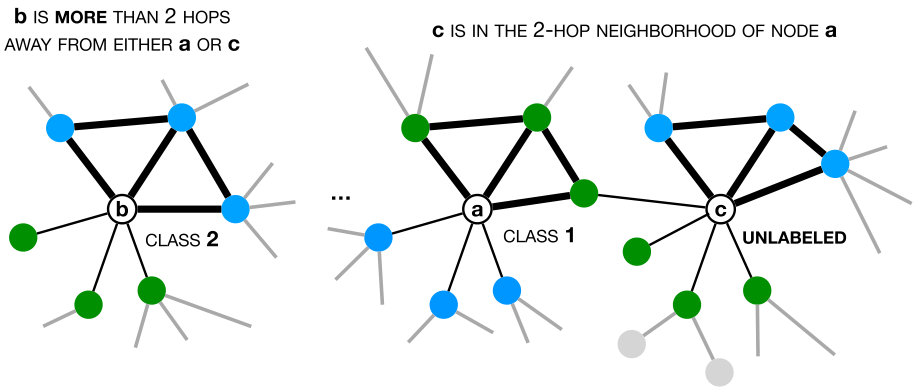


Fig. 1 Localized message passing limitations: A stylized example in a homogeneous graph with a 2-layer GNN (colors indicate node attributes). Node **a** is in the 2-hop range of node **c**. Node **c** does not receive gradient updates from node **b** (class 2) since node **b** is more than 2 hops away.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

its graph neighbors via learnable aggregators. Long-range dependencies are captured by using k layers to incorporate features from k-hop neighborhoods. GNNs have demonstrated promising results in several application domains spanning *homogeneous* graphs (*e.g.*, user-user friendship networks) comprising nodes and edges of a single type, and *heterogeneous* graphs (*e.g.*, academic citation networks) containing nodes and edges of different types.

Localized message passing limitations We illustrate two key limitations of prior k-layer GNN architectures: *k-hop localized* and *over-smoothed* representations (Fig. 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 Fig. 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 [25]. In Fig. 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 covariation difference when classifying node c.

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

One way to overcome these limitations is the paradigm of role discovery [38] 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 [41]; 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 [36], *i.e.*, they will assign the same role to nodes a and b in Fig. 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. Furthermore, structural role learning is relatively unexplored in heterogeneous graphs with typed nodes and edges.

Present Work To enable the expressivity to distinguish attributed structures, we propose the concept of *attributed structural roles* that identify structurally similar nodes with covarying attributes, independent of network proximity. We ground structural roles on *network motifs*,[1] which are induced subgraph structures over a few nodes (*e.g.*, triangles). Networks motifs are a broad class of *higher-order structures* that indicate connectivity patterns between nodes, and are crucial for understanding the organization and properties of complex networks [28]. In addition, network motifs can be easily generalized to capture type semantics in rich heterogeneous graphs through heterogeneous (typed) higher-order structures [39]. Leveraging higher-order connectivity structures between nodes is extremely valuable to overcome the lack of sufficient training labels in local neighborhoods during semi-supervised learning. We define two nodes as sharing attributed structural roles if they participate in topologically similar motif instances over covarying sets of attributes. We note that attribute covariance permits for multiple discrete and continuous attributes, rather than stricter notions such as regular equivalence [41].

We propose InfoMotif, a GNN architecture-agnostic regularization framework that exploits the covariance 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 are:

- *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 [36], we adopt the paradigm of *self-supervised learning* to regularize node representations to capture *attribute correlations* in motif structures. InfoMotif unifies the expressive local neighborhood aggregation power of message-passing GNNs with the paradigm of structural roles.
- *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 [61, 64], we achieve architecture independence by modulating the node representations learned by the base GNN through motif-based *mutual information maximization*, to capture attributed structural roles. We regularize three state-of-the-art GNNs within our framework to demonstrate considerable performance gains over several prior graph learning approaches.
- *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 [70, 72] 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.

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

We regularize three state-of-the-art GNN models in our InfoMotif framework for semi-supervised node classification. Our experiments are conducted on a wide variety of real-world datasets spanning homogeneous and heterogeneous networks. In *homogeneous* graphs, InfoMotif outperforms prior approaches (by 3-10% classification accuracy) on two diverse classes of datasets: assortative *citation* networks that exhibit strong homophily and dis-assortative *air-traffic* networks that depend on structural roles. We also demonstrate the utility of our framework in three *heterogeneous* graph datasets where InfoMotif outperforms a number of state-of-the-art methods with notable performance gains (5% accuracy) on average. 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 Sect. 3, we present the problem formulation, and introduce preliminaries on GNNs and network motifs. We describe our proposed framework InfoMotif in Sect. 4 with architectural details in Sect. 5, present experimental results in Sect. 6, discussions in Sect. 7, and finally conclude in Sect. 8.

2 Related work

Our work is related to semi-supervised learning approaches over *homogeneous* and *heterogeneous* graphs, and recent advances in the paradigm of *self-supervised* learning.

Homogeneous Graphs Semi-supervised learning over graphs is a well-studied problem, where the goal is to classify nodes in a graph given a small set of labeled examples. The most popular label spreading [70, 71] techniques propagate labels through linked nodes in the graph based on smoothness assumptions. Graph neural networks (GNNs) generalize label spreading through localized message passing over feature-rich node neighborhoods and have achieved state-of-the-art results in several benchmarks [22]. GNNs learn node representations by recursively aggregating features from local neighborhoods in an end-to-end manner, with diverse applications, including information diffusion prediction [50], friend suggestions [44], social recommendation [23], and community question answering [29]. Graph convolutional networks (GCNs) [22] learn degree-weighted aggregators, which can be interpreted as a special form of Laplacian smoothing [25]. Many models generalize GCN with a wide range of neighborhood aggregators, *e.g.*, self-attentions [46, 57], mean and max pooling functions [15], 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 prior graph representation learning approaches that aim to overcome the key limitations of *oversmoothing* and *localization* in GNNs: *non-local GNNs* that capture contributions from distant nodes in the graph and *structural role learning* techniques that enhance the structural distinguishability of the learned node representations.

Non-local methods expand the propagation range of GNNs to aggregate node representations of differing localities, *e.g.*, JKNet [64] uses skip-connections to vary the influence radius per node, PGNN [65] captures global network positions via shortest-paths, and DGI [58] maximizes MI between node representations and a summary representation of the entire graph. However, they either operate on a local scale [33] 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 [17, 41]. A few approaches [55] 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 statis-

tics) to extend classical proximity-preserving embedding methods, *e.g.*, feature-based matrix factorization [40] and random walk methods [36]. Notably, a few methods design structural GCNs via motif adjacency matrices [24, 48, 49]. However, all these methods model structural roles without considering node attributes. A related direction is higher-order network representation learning that models proximity via network motifs [8]. However, such representations are highly localized and cannot identify structurally similar nodes independent of network proximity. In contrast, we regularize GNNs to learn attributed structural roles based on the covariance of attributes in motifs, thus simultaneously enhancing the distinguishability of node representations and identifying correspondences between distant nodes.

Heterogeneous Graphs Representation learning techniques over heterogeneous graphs primarily focus on preserving structural information indicated by the type semantics in meta-path or meta-graph structures. A few popular approaches include node representation learning by capturing proximities between node pairs connected via meta-graphs [53, 68] and meta-path guided random walks [10, 12]. Recently, graph neural networks have been generalized to heterogeneous graphs through message passing aggregation over local neighborhoods induced via specific node types [19, 67], meta-paths [13, 60, 69] and meta-graphs [49]. While these advances effectively incorporate rich heterogeneous semantics into message-passing GNNs, the key limitation due to localization remains. To our knowledge, structure role learning in heterogeneous graphs is unexplored and ours is the first to examine structural role learning in GNNs with rich type semantics and attributes.

Self-supervised Learning The emerging paradigm of self-supervised learning [18] aims to alleviate the need for large volumes of labeled examples by extracting supervision signals from the intrinsic structure of the raw data. For instance, auxiliary supervision signals for images are created by rotating, cropping and colorizing images, followed by new training objectives to facilitate representation learning [5]. One empirically effective strategy is mutual information maximization [3] to maximize agreement across different views of the data. A few recent advances extend self-supervised learning [43, 45] to graph representation learning by exploiting structural properties such as node degree, proximity [32], and attributes [21], for model pre-training [20, 26, 34, 63, 66]. In our work, we design self-supervised learning objectives to regularize graph neural networks for node classification by learning attribute correlations in higher-order connectivity patterns (typed and untyped motif structures).

3 Preliminaries

In this section, we formalize semi-supervised node classification on graphs via graph neural networks and introduce network motifs in homogeneous and heterogeneous graphs.

3.1 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 $v \in \mathcal{V}$. In our work, the graph may be *heterogeneous* with multiple types of nodes and edges; in such a scenario, we have a node type mapping $\psi : \mathcal{V} \mapsto \mathcal{T}_V$ where $\mathcal{T}_V$ is the set of node types that identifies each node in $\mathcal{V}$ with a type in $\mathcal{T}_V$, and a corresponding edge type mapping $\xi : \mathcal{E} \mapsto \mathcal{T}_E$ where $\mathcal{T}_E$ is the set of edge types. Each labeled node $v \in \mathcal{V}_L$ belongs

to one of C classes, encoded by a one-hot vector $\mathbf{y}_v \in \mathbb{B}^C$ ($\mathbb{B} = \{0, 1\}$) where C denotes the number of classes. Our goal is to predict the labels of the unlabeled nodes $v \in \mathcal{V}_U$. This is the familiar transductive or semi-supervised learning setup for node classification in graphs [70].

3.2 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 nonlinear 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 [22], graph attention [57], and pooling [15]. Let F denote the cardinality of node features at the input layer, and D indicate the embedding size after the final GNN layer. Thus, 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 [6] that learn convolutional filters from the entire graph.

3.3 Network motifs

Network motifs are a general class of higher-order connectivity patterns, with a history of use in network science [28, 30]. A motif has several topologically equivalent appearances in the network called *motif instances*. Prior work [37, 42] 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 denote the number of nodes in network motif M_t; that is, $k_t = |\mathcal{V}_t|$.

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

In a heterogeneous graph, nodes/edges are of many different types which makes it essential to explicitly (and jointly) model the connectivity patterns and the participating types. We define *typed* network motifs that generalize network motifs through additional constraints on the types of participating nodes, below:

Definition 2 (*Typed Network Motif*). A typed network motif denoted by $M_t = (\mathcal{V}_t, \mathcal{E}_t, \psi_t, \xi_t)$ is a connected, induced subgraph consisting of a subset $\mathcal{V}_t \subset \mathcal{V}$ and $\mathcal{E}_t = \{e \in \mathcal{E} \mid e = (u, v), u, v \in \mathcal{V}_t\}$ such that node and edge type mappings $\psi_t = \psi|_{\mathcal{E}_t}$ and $\xi_t|_{\mathcal{V}_t}$ are restrictions of ψ and ξ to $\mathcal{V}_t$ and $\mathcal{E}_t$, respectively, and $k_t = |\mathcal{V}_t|$ is the number of nodes in M_t.

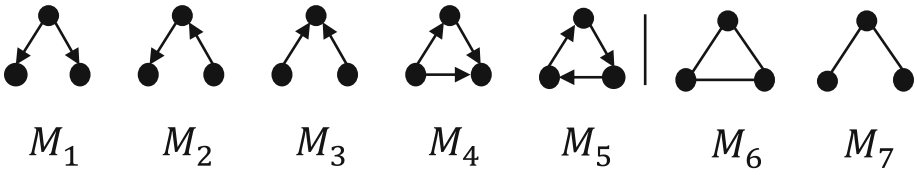
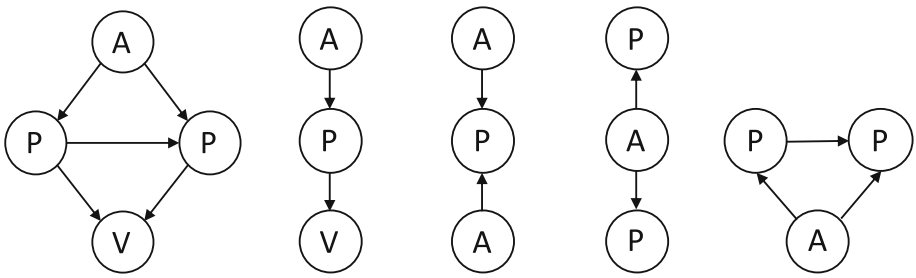


Fig. 2 Topologically distinct, directed (M_1 to M_5) and undirected (M_6 to M_7) 3-node, connected, network motifs



(a) Schema of DBLP **(b)** Examples of typed 3-node motifs in DBLP

Fig. 3 **a** Heterogeneous network schema of bibliographic network DBLP with three node types: author (A), paper (P), and venue (V) and three edge types $A \rightarrow P$, $P \rightarrow V$, and $P \rightarrow P$. **b** Examples of 3-node connected typed network motifs

We assume that the given graph $\mathcal{G}$ has a set of unique associated network motifs $\mathcal{M} = \{M_1, \ldots, M_T\}$.

Definition 3 (*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 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 $\mathcal{I}_v(M_t)$.

3.4 Model 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 $\mathbf{y_v} \in \mathbb{B}^{\mathbf{C}}$ of training nodes in $v \in \mathcal{V}_L$, defined by:

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

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 [31]) 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.

4 InfoMotif framework

In this section, we first discuss the structural properties of GNNs to motivate the notion of attributed structural roles. In Sect. 4.2, we present our motif-based self-supervised learning framework InfoMotif to regularize GNNs based on a single motif. Finally, in Sect. 4.3, we introduce our overall framework with a novel multi-motif training curriculum.

4.1 Motivating insights: attributed structural roles

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 $\boldsymbol{X}(S)$ denote its set of input node features. Let $\boldsymbol{Y}(\mathcal{V}_L)$ 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 (Eq. 2), let $\Theta^* = \{\Theta_1, \ldots, \Theta_k\}$ be the optimal parameters computed by its training algorithm. Now, we have the following proposition.

Proposition 1 *Θ^* is a function of $\boldsymbol{X}(\mathcal{N}_k(\mathcal{V}_L))$, $\boldsymbol{Y}(\mathcal{V}_L)$ and changes in inputs $\boldsymbol{X}(\mathcal{V} \setminus \mathcal{N}_k(\mathcal{V}_L))$ will not affect Θ^*.*

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

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 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 1:

Corollary 1 *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 Fig. 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 covariation of attributes and structure (blue neighbors form triangles), node c does not influence the training loss for all nodes with class 2 (Table 1).

4.2 Self-supervised single motif regularization

In this section, we introduce InfoMotif, a framework to regularize node representations of the base GNN by exploiting the covariance of node attributes and motif structures. We define *attributed structural roles* by assigning the same role to nodes that participate in motif instances over *covarying* sets of attributes. Compared to prior role-aware models [36] that discover structurally similar nodes *agnostic* to attributes, we define roles based on attribute

Table 1 Notation

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 $v \in \mathcal{V}_L$

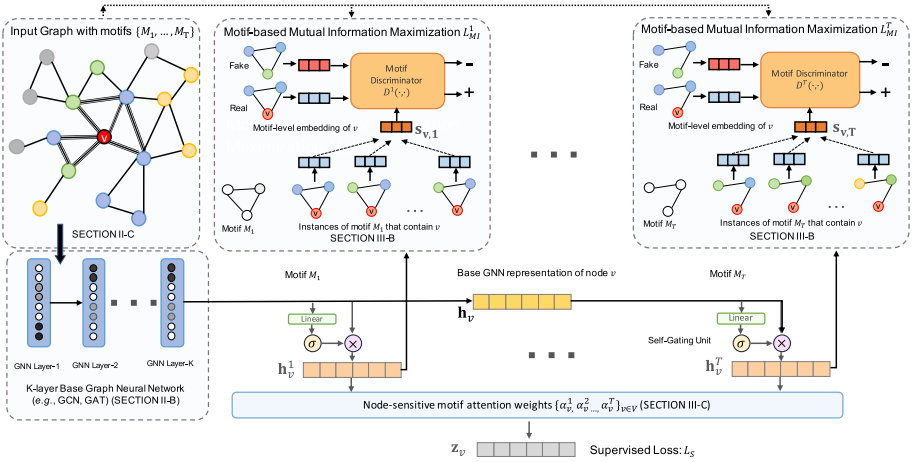


Fig. 4 Architecture diagram of InfoMotif depicting the model components: base GNN $f_{\theta,l}$ with k layers (bottom left), motif-based mutual information maximizing regularizers L_{MI}^t (top right), and attention module to compute final node representations $\mathbf{z}_v$ (bottom right). Instances of motif M_1 are shown in the graph (top left) with textured lines and colors indicate node attributes

occurrence in higher-order connectivity structures. In heterogeneous graphs, roles further incorporate semantics of node and edge types described by the connectivity structures of typed network motifs.

Now, we describe our self-supervised learning strategy to learn attribute covariance 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 covariance 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. (Detailed descriptions are presented in Sects. 5.2 and 5.3.) By introducing motif-based MI maximization as a regularizer, the GNN is encouraged to learn *discriminative* statistical correspondences between nodes that participate in instances of the same motif. Motif-based MI maximization is an example of the broader paradigm of self-supervised learning that derives auxiliary supervision signals from the intrinsic structure (*e.g.,* connectivity patterns in a network motif) of the underlying data.

We first adapt the base GNN representation $\mathbf{h}_v$ (see Sect. 3.2), specific to motif M_t through a *motif gating* function $f^t_{\text{GATE}} : \mathbb{R}^D \mapsto \mathbb{R}^D$ resulting in a gated embedding $\mathbf{h}^t_v$. Then, we introduce a *motif instance encoder* $f^t_{\text{ENC}} : \mathbb{R}^D \times \mathbb{R}^{(k_t \times D)} :\mapsto \mathbb{R}^D$ 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 $f^t_{\text{READ}}(\cdot)$, *e.g.*, averaging or pooling functions.

For each node $v \in \mathcal{V}$, we maximize motif-based mutual information L^t_{MI} 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:

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

where θ and ϕ^t denote the parameters of the layers $\{f_{\theta,l}\}^k_{l=1}$, and motif-specific transforms $\{f^t_{\text{GATE}}, f^t_{\text{ENC}}, f^t_{\text{READ}}\}$, respectively. By maximizing MI across all instances of motif M_t in the graph through a shared MI estimator I^t_ψ, 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 [4, 18], 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 *contrastive* discriminator network $\boldsymbol{D}^t_\psi : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}^+$, where $\boldsymbol{D}^t_\psi(\mathbf{e}_{v,I_t}, \mathbf{s}_{v,t})$ denotes the probability score assigned to this instance-motif pair. The positive samples $(\mathbf{e}_{v,I_t}, \mathbf{s}_{v,t})$ for $\boldsymbol{D}^t_\psi$ 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 $\boldsymbol{D}^t_\psi$ is trained on a noise-contrastive objective L^t_{MI} between samples from the joint (positive pairs), and the product of marginals (negative pairs), which is defined as:

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

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 [58].

We design the negative sampling distribution $P_{\mathcal{N}}(\widetilde{I}_t|M_t)$ to learn attribute covariance 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, $\boldsymbol{D}^t_\psi$ learns attributed structural roles with respect to motif M_t.

4.3 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 (Eq. 4) from multiple motifs is given by:

$$L^{'} = L_B + \lambda L^{'}_{MI} = L_B + \lambda \cdot \frac{1}{T} \sum_{t=1}^{T} L^t_{MI} \tag{5}$$

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 toward 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 [2], 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 (Eq. 2) over the labeled node set $\mathcal{V}_L$. We instead aggregate the set of motif-gated representations ($\mathbf{h}^t_v$ 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:

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

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 (Eq. 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 (Eq. 5), we contextually weight the contributions of different motif regularization terms (Eq. 4) at a node-level granularity through the attention weights α_{vt} of motif M_t for node v.

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

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 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.

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

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 [35]. In particular, re-weighting strategies that emphasize harder examples are effective at handling imbalanced data distributions [7]. 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 $v \in \mathcal{V}_L$.

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

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

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

In contrast to the original supervised loss L_B (Eq. 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 (Eq. 9), and motif regularizers (Eq. 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 hyper-parameter 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 is 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.

5 Model details

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

5.1 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 $f^t_{\text{GATE}}(\cdot)$ for motif M_t learns a nonlinear gate to modulate the input at a feature-wise granularity through dimension re-weighting, defined by:

$$\mathbf{h}^t_v = f^t_{\text{GATE}}(\mathbf{h}_v) = \mathbf{h}_v \odot \sigma(\mathbf{W}^t_g \mathbf{h}_v + \mathbf{b}^t_g) \tag{11}$$

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 nonlinearity. The self-gating mechanism effectively serves as a multiplicative skip-connection [9] that facilitates gradient flow from the motif-based regularizer to the GNN.

5.2 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 [56] 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, $\mathbf{h}_v^t$.

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

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.

5.3 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:

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

where σ is the sigmoid nonlinearity. We adopt batch-wise training with motif instance sampling ($\sim$ 20 per node) to compute $\mathbf{s}_{v,t}$. Sophisticated readout architectures [59] are more likely necessary to handle larger sample sizes.

5.4 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 [47, 58], 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 nonlinearity to convert raw scores into probabilities of $(\mathbf{e}_{v,I_t}, \mathbf{s}_v^t)$ being a positive example for motif M_t.

6 Experiments

We present extensive quantitative and qualitative analyses on multiple diverse datasets spanning homogeneous and heterogeneous graphs. We first introduce datasets, baselines, and experimental setup (Sects. 6.1, 6.2, and 6.3). We propose four research questions to guide our experiments:

(RQ_1) How does InfoMotif compare with state-of-the-art graph neural networks and embedding learning methods on *node classification* over *homogeneous* graphs?

Table 2 Dataset statistics of homogeneous network benchmarks, including three assortative citation [51] networks and three dis-assortative air-traffic [36] networks

	Citation networks			Air-traffic networks		
Dataset	Cora	Citeseer	PubMed	Brazil	Europe	USA
# Nodes	2485	2110	19,717	131	399	1190
# Edges	5069	3668	44,324	1038	5995	13,599
# Attributes	1433	3703	500	–	–	–
# Classes	7	6	3	4	4	4

Ground-truth classes in citation networks exhibit attribute homophily; ground-truth classes in flight networks indicate node structural roles

Table 3 Dataset statistics of three heterogeneous graphs across bibliographic and movie networks, with multiple node and edge types

Dataset	DBLP-A	DBLP-P	Movie
# Nodes	11,170	35,770	10,441
# Edges	24,846	131,636	99,509
# Attributes	4479	11,680	4577
# Node Types	3	3	4
# Classes	4	10	6

Schemas are shown in Figs. 3a and 5a

(RQ_2) Is InfoMotif effective for *node classification* on *heterogeneous* information networks (multiple types of nodes and edges) compared to state-of-the-art approaches?
(RQ_3) How do the different *architectural* design choices and *training* strategies in InfoMotif impact performance?
(RQ_4) What is the impact of *node degree*, local training *label sparsity*, and local *attribute diversity*, on the classification performance of InfoMotif?
(RQ_5) How do the motif-based *regularization* strategies and associated *hyper-parameters* in InfoMotif affect model training time and performance?

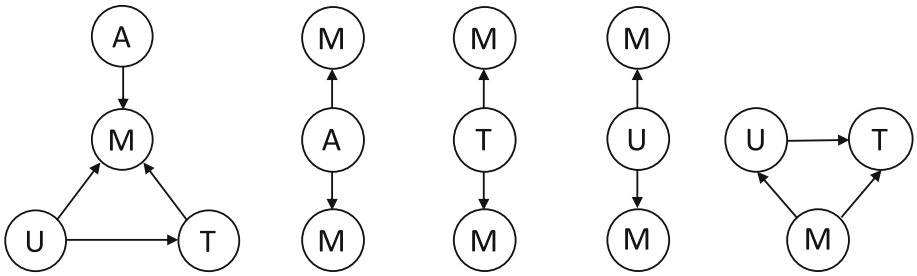
6.1 Datasets

Our experiments are designed toward semi-supervised node classification on nine real-world benchmark datasets, divided across homogeneous and heterogeneous networks.

In homogeneous graphs, we experiment on two diverse types of datasets: *citation networks* that exhibit strong homophily and *air-traffic networks* that depend on structural roles (Table 2).

– *Citation Networks* We consider three benchmark datasets, Cora, Citeseer, and PubMed [51], 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 the USA [36] 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.

We conduct experiments on three real-world datasets over heterogeneous graphs whose statistics are shown in Table 3:



(a) Schema of Movie **(b)** Examples of typed 3-node motifs in MovieLens

Fig. 5 **a** Heterogeneous network schema of movie network with four node types: actor (A), movie (M), user (U), and term (T) and four edge types $A \rightarrow M, U \rightarrow M, T \rightarrow M$, and $U \rightarrow T$. **b** Examples of 3-node connected typed network motifs

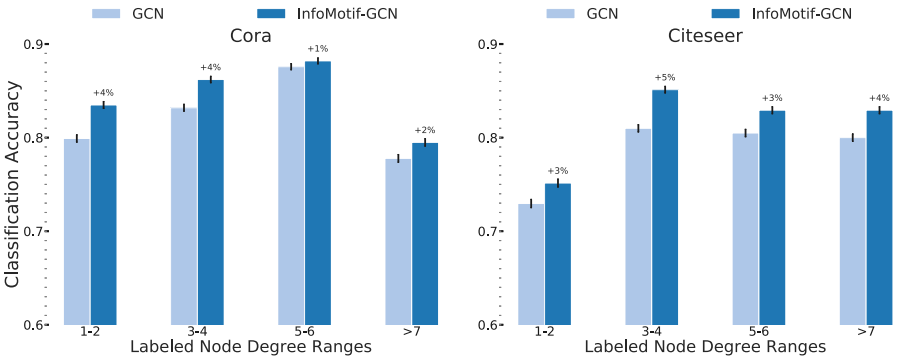


Fig. 6 Classification accuracy with respect to node degree. InfoMotif has consistent gains across all segments with higher gains for low-to-medium degree nodes (quartiles Q1 and Q2)

– *DBLP-A* This is a bibliographic network composed of 3 node types: author (A), paper (P), and venue (V), connected by three link types: $P \rightarrow P$, $A \rightarrow P$, and $P \rightarrow V$ (Fig. 3a). We use a subset of DBLP [54] with text attributes of papers to classify authors based on their research areas.
– *DBLP-P* This dataset has the same schema as DBLP-A, but the learning task is to classify research papers into ten categories, which are obtained from Cora [27].
– *Movie* We use MovieLens [16] with four node types: movie (M), user (U), actor (A), and tag (T) linked by four types: $U \rightarrow M$, $A \rightarrow M$, $U \rightarrow T$, and $T \rightarrow M$, with attributes for actors and movies. The classification task is movie genre prediction (Fig. 4).

6.2 Baselines

We first introduce baseline methods designed for learning over homogeneous graphs, organized into four categories based on whether they are *proximity-based vs. structural*, and the paradigm of *embedding learning vs. graph neural networks*:

– *Proximity-based embedding methods* Conventional methods, node2vec [14] that learns from second-order random walks, and motif2vec [8] that models higher-order proximity.

– *Structural embedding methods* Structural role-aware models struc2vec [36], GraphWave [11], and DRNE [55].
– *Standard Graph Neural Networks* State-of-the-art GNN models based on localized message passing: GCN [22], GraphSAGE [15], GAT [57], JKNet [64], and DGI [58].
– *Structural Graph Neural Networks* Motif-based Motif-CNN [48], MCN [24], and degree-specific DEMO-Net [61].

Next, we introduce graph neural networks designed specifically for semi-supervised learning over heterogeneous graphs.

– *Heterogeneous Graph Neural Networks* State-of-the-art *metapath*-based GNN models MAGNN [13], HAN [60], and *metagraph*-aware GNN model Meta-GNN [49].

6.3 Experimental setup

We tested our framework InfoMotif by integrating GCN, JKNet, and GAT as base graph neural networks for motif-based regularization. We only consider the largest connected component in each dataset and evaluate different train/validation/test splits to fairly compare different models [52]. We create 10 random data splits per training ratio and evaluate 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] We train two-layer base GNNs for GCN and GAT (8 attention heads per layer) models while training the base JKNet using 4 GCN layers and maxpool layer aggregation. The model is trained for a maximum of 100 epochs with a batch size of 256 nodes with Adam optimizer. We also apply dropout for model regularization with a rate of 0.5 and tune the learning rate in the range $\{10^{-4}, 10^{-3}, 10^{-2}\}$.

6.4 Homogeneous graphs (RQ_1)

In homogeneous graphs, we train InfoMotif using the set of all directed 3-node motifs in citation networks and undirected 3-node motifs in air-traffic networks (Fig. 2). For citation networks, we train base GNNs with layer sizes of 256 each, while using 64 for the smaller air-traffic networks. We evaluate different train/validation/test splits (training ratios of 20% and 40%) and report experimental results comparing InfoMotif with three base GNNs, against competing baselines on citation and air-traffic networks, in Tables 4 and 5, 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. JKNet 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.

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

Table 4 Node classification results (% test accuracy) on assortative citation networks using 10 random train/validation/test splits per training ratio (20% and 40%)

Training ratio	Data		Cora		Citeseer		PubMed	
	X	Y	20%	40%	20%	40%	20%	40%
Proximity-based graph embedding methods								
Node2Vec [14]			75.7 ± 0.5	76.1 ± 0.5	68.1 ± 0.5	69.1 ± 0.6	80.1 ± 0.6	80.2 ± 0.6
Motif2Vec [8]			79.0 ± 0.4	79.2 ± 0.4	66.6 ± 0.4	67.1 ± 0.3	79.8 ± 0.2	79.8 ± 0.4
Structural graph embedding methods								
Struct2Vec [36]			35.4 ± 1.0	37.6 ± 1.3	31.2 ± 0.8	35.1 ± 0.9	48.5 ± 0.3	49.2 ± 0.4
GraphWave [11]			39.5 ± 2.1	41.1 ± 1.5	38.5 ± 1.2	40.6 ± 0.9	43.0 ± 2.0	43.3 ± 1.3
DRNE [55]			34.9 ± 1.5	36.5 ± 1.5	30.8 ± 1.2	32.2 ± 1.2	40.4 ± 0.7	41.6 ± 0.4
Standard graph neural networks								
GCN [22]	✓	✓	81.6 ± 0.5	82.0 ± 0.4	75.8 ± 0.5	76.6 ± 0.3	85.7 ± 0.7	86.1 ± 0.5
GAT [57]	✓	✓	80.9 ± 0.7	81.4 ± 0.2	74.5 ± 0.7	75.5 ± 0.7	83.3 ± 0.3	84.2 ± 0.3
GraphSAGE [15]	✓	✓	81.3 ± 0.3	83.5 ± 0.3	72.9 ± 0.3	73.8 ± 0.2	86.6 ± 0.2	87.2 ± 0.3
JKNet [64]	✓	✓	81.3 ± 0.8	83.6 ± 0.8	71.5 ± 0.8	72.5 ± 0.7	82.2 ± 0.4	83.8 ± 0.5
DGI [58]	✓		76.2 ± 0.8	77.3 ± 0.9	74.5 ± 0.7	74.7 ± 0.7	78.2 ± 0.9	78.5 ± 0.9
Structural graph neural networks								
DemoNet [61]	✓	✓	81.0 ± 0.6	82.4 ± 0.5	67.9 ± 0.7	68.5 ± 0.6	79.5 ± 0.4	80.5 ± 0.4
Motif-CNN [48]	✓	✓	81.6 ± 0.5	82.8 ± 0.5	73.4 ± 0.3	76.8 ± 0.3	87.3 ± 0.1	87.5 ± 0.1
MCN [24]	✓	✓	81.1 ± 0.9	82.4 ± 0.8	73.2 ± 0.4	75.9 ± 0.7	85.2 ± 0.6	85.9 ± 0.5
Motif-regularized graph neural networks (InfoMotif)								
InfoMotif-GCN	✓	✓	**85.7 ± 0.4**	**87.4 ± 0.4**	**77.7 ± 0.5**	**78.5 ± 0.5**	**87.5 ± 0.2**	**88.3 ± 0.2**
InfoMotif-JKNet	✓	✓	85.5 ± 0.3	86.5 ± 0.5	74.5 ± 0.8	76.7 ± 0.9	87.0 ± 0.2	87.9 ± 0.3
InfoMotif-GAT	✓	✓	85.5 ± 0.3	87.2 ± 0.7	76.5 ± 0.5	77.0 ± 0.4	85.9 ± 0.4	86.2 ± 0.5

X and **Y** denote the use of node attributes and training labels, respectively, toward 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
The bold font indicates the accuracy numbers of our model InfoMotif (and variants) with highest performance

Table 5 Node classification results (% test accuracy) on dis-assortative air-traffic networks

Training ratio	Data		USA		Europe		Brazil	
	X	Y	20%	40%	20%	40%	20%	40%
Proximity-based graph embedding methods								
Node2Vec [14]			24.6 ± 0.9	24.8 ± 0.9	36.5 ± 1.0	37.4 ± 1.1	26.3 ± 1.4	30.4 ± 1.3
Motif2Vec [8]			51.3 ± 1.1	54.8 ± 1.1	37.1 ± 1.2	38.1 ± 1.2	27.2 ± 1.5	33.9 ± 1.5
Structural graph embedding methods								
Struct2Vec [36]			50.4 ± 0.8	51.3 ± 0.8	42.5 ± 0.7	45.6 ± 0.8	45.8 ± 1.1	51.8 ± 1.1
GraphWave [11]			45.2 ± 1.4	48.0 ± 1.4	38.1 ± 1.9	41.1 ± 1.6	40.2 ± 2.0	43.1 ± 1.8
DRNE [55]			51.3 ± 1.1	52.4 ± 1.1	43.1 ± 1.7	47.6 ± 1.3	46.5 ± 2.7	50.2 ± 2.3
Standard graph neural networks								
GCN [22]	✓	✓	51.9 ± 0.9	56.0 ± 0.9	37.4 ± 0.9	40.1 ± 0.8	36.5 ± 1.5	38.9 ± 1.6
GAT [57]	✓	✓	52.7 ± 1.0	53.5 ± 0.9	31.5 ± 1.0	34.3 ± 1.0	37.3 ± 1.6	37.9 ± 1.6
GraphSAGE [15]	✓	✓	45.3 ± 1.2	49.4 ± 1.2	28.8 ± 1.0	32.5 ± 1.0	36.1 ± 1.6	37.5 ± 1.6
JKNet [64]	✓	✓	53.8 ± 1.2	56.1 ± 1.0	49.7 ± 1.1	53.8 ± 1.1	55.9 ± 1.5	58.4 ± 1.8
DGI [58]	✓		46.4 ± 1.3	47.3 ± 1.2	37.5 ± 1.5	39.9 ± 1.5	41.4 ± 1.6	45.2 ± 1.7
Structural graph neural networks								
DemoNet [61]	✓	✓	58.6 ± 1.2	58.8 ± 1.1	40.4 ± 1.3	46.2 ± 1.2	46.1 ± 1.4	48.9 ± 1.5
Motif-CNN [48]	✓	✓	53.6 ± 1.0	54.2 ± 1.0	37.9 ± 1.0	41.1 ± 1.1	28.9 ± 1.6	35.7 ± 1.7
MCN [24]	✓	✓	54.8 ± 1.4	54.9 ± 1.3	36.8 ± 1.2	39.6 ± 1.5	42.9 ± 1.6	43.6 ± 1.4
Motif-regularized graph neural networks (**InfoMotif**)								
InfoMotif-GCN	✓	✓	59.5 ± 0.9	62.9 ± 0.7	**53.5 ± 0.6**	56.9 ± 0.6	56.6 ± 1.2	60.7 ± 1.2
InfoMotif-JKNet	✓	✓	**61.8 ± 1.6**	**64.3 ± 1.2**	53.1 ± 1.2	**56.9 ± 0.6**	**62.7 ± 1.8**	**67.9 ± 1.5**
InfoMotif-GAT	✓	✓	58.0 ± 0.4	60.4 ± 0.3	46.0 ± 1.5	50.0 ± 2.0	50.6 ± 1.3	56.3 ± 1.1

Structural embedding methods and GNNs outperform proximity-based models. InfoMotif JKNet achieves significant gains of 4–14% across datasets
The bold font indicates the accuracy numbers of our model InfoMotif (and variants) with highest performance

Table 6 Node classification results (% test accuracy) on heterogeneous graphs from bibliographic and movie networks

Training ratio	Data		DBLP-A		DBLP-P		Movie	
	X	Y	10%	20%	10%	20%	10%	20%
Proximity-based graph embedding methods								
Node2Vec [14]			63.9 ± 0.4	65.3 ± 0.6	68.5 ± 0.4	70.1 ± 0.5	54.1 ± 0.3	56.7 ± 0.3
Motif2Vec [8]			62.7 ± 0.7	64.5 ± 0.6	68.9 ± 0.9	71.3 ± 1.1	55.0 ± 0.3	57.4 ± 0.5
Structural graph embedding methods								
Struct2Vec [36]			34.2 ± 0.4	36.1 ± 0.4	34.9 ± 0.2	35.7 ± 0.3	32.7 ± 0.4	34.9 ± 0.1
GraphWave [11]			34.8 ± 0.5	37.0 ± 0.6	35.6 ± 0.2	36.3 ± 0.3	33.5 ± 0.4	36.0 ± 0.4
DRNE [55]			33.9 ± 0.3	36.5 ± 0.3	35.1 ± 0.5	35.5 ± 0.4	31.0 ± 0.2	35.1 ± 0.2
Standard graph neural networks								
GCN [22]	✓	✓	65.3 ± 1.1	69.6 ± 0.9	71.3 ± 0.7	73.4 ± 0.8	55.7 ± 1.0	57.3 ± 0.7
GAT [57]	✓	✓	67.5 ± 0.8	71.7 ± 0.8	71.9 ± 0.5	73.0 ± 0.7	58.6 ± 0.9	59.9 ± 1.0
GraphSAGE [15]	✓	✓	65.3 ± 0.7	69.0 ± 0.6	70.9 ± 0.8	72.7 ± 0.8	55.6 ± 0.4	56.4 ± 0.8
JKNet [64]	✓	✓	69.6 ± 1.0	73.2 ± 1.2	69.8 ± 1.1	72.0 ± 1.3	58.3 ± 0.7	60.5 ± 0.9
DGI [58]	✓		64.7 ± 0.5	68.5 ± 0.7	41.9 ± 0.8	61.1 ± 0.6	38.6 ± 0.9	40.4 ± 0.9
Structural graph neural networks								
DemoNet [61]	✓	✓	70.7 ± 1.3	72.3 ± 1.1	72.6 ± 0.9	73.5 ± 0.6	59.5 ± 0.8	61.2 ± 0.8
Motif-CNN [48]	✓	✓	66.4 ± 1.1	70.1 ± 1.3	71.5 ± 0.8	72.2 ± 0.6	54.3 ± 0.3	56.9 ± 0.5
MCN [24]	✓	✓	67.1 ± 1.2	71.2 ± 1.1	71.9 ± 0.9	72.5 ± 0.6	54.7 ± 0.4	57.2 ± 0.6
Heterogeneous graph neural networks								
HAN [60]	✓	✓	68.2 ± 1.0	72.0 ± 1.3	73.1 ± 0.9	74.0 ± 0.7	60.7 ± 1.1	62.1 ± 0.8
MAGNN [13]	✓	✓	68.9 ± 0.7	72.5 ± 0.7	74.7 ± 0.6	75.8 ± 0.7	62.1 ± 0.9	63.0 ± 0.5
Meta-GNN [49]	✓	✓	71.3 ± 1.2	73.9 ± 1.4	74.6 ± 0.6	75.8 ± 0.6	61.7 ± 0.5	63.7 ± 0.7

Table 6 continued

Training ratio	Data		DBLP-A		DBLP-P		Movie	
	X	Y	10%	20%	10%	20%	10%	20%
Motif-regularized graph neural networks (InfoMotif)								
InfoMotif-GCN	✓	✓	73.7 ± 1.2	77.4 ± 1.1	**78.8 ± 0.4**	**79.0 ± 0.7**	**64.7 ± 0.8**	**65.0 ± 1.1**
InfoMotif-JKNet	✓	✓	**75.6 ± 0.9**	**79.9 ± 0.8**	75.5 ± 1.0	76.3 ± 1.2	60.7 ± 0.6	62.0 ± 0.6
InfoMotif-GAT	✓	✓	72.4 ± 1.0	75.3 ± 1.4	77.1 ± 0.7	77.9 ± 0.6	62.8 ± 0.5	64.2 ± 0.5

Heterogeneous GNN models (MAGNN, Meta-GNN) typically outperform structural GNNs (DemoNet, Motif-CNN, MCN) and type-agnostic message-passing GNNs (GCN, GAT). InfoMotif achieves significant gains of 5% on average across datasets

The bold font indicates the accuracy numbers of our model InfoMotif (and variants) with highest performance

Table 7 Ablation study results with 40% training ratio on citation networks

Dataset	Cora	Citeseer	PubMed
InfoMotif-GCN ($L_S + \lambda L_{MI}$)	**87.4 ± 0.4**	**78.5 ± 0.5**	**88.3 ± 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

The novelty and task weighting strategies improve classification accuracies by 2% on average
The bold font indicates the accuracy numbers of our model InfoMotif (and variants) with highest performance

6.5 Heterogeneous graphs (RQ_2)

In heterogeneous graphs, we train InfoMotif and other baselines that use motifs/metagraphs using typed 3-node network motifs (shown in Figs. 3b, 5b) that are defined based on the heterogeneous type schema (Figs. 3a, 5a). Our experimental results comparing our framework InfoMotif with baselines on DBLP and movie networks are shown in Table 6.

We find that message-passing GNNs (such as GCN and GAT) generally outperform conventional embedding methods (such as node2vec). Structural embedding methods (*e.g.*, struc2vec) perform poorly; this reveals their inability to capture structural aspects relevant to heterogeneous graphs with rich type semantics. Heterogeneous GNNs such as MAGNN and Meta-GNN outperform homogeneous GNNs owing to their type-aware semantic neighbor aggregation via metapaths and metagraphs, respectively. Our framework InfoMotif further learns type-aware attributed structural roles which results in significant performance gains of 5% on average over prior approaches.

6.6 Model ablation study (RQ_3)

We present an ablation study on citation networks to analyze the importance of major components in InfoMotif (Table 7)

- *Remove novelty-driven sample weighting* We set the novelty $\beta_v = 1$ (Eq. 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 (Eq. 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.

6.7 Qualitative analysis (RQ_4)

We qualitatively examine the source of InfoMotif's gains over the base GNN (GCN due to its consistent performance), by analyzing *node degree*, *label sparsity*, and *attribute diversity* in local node neighborhoods, on the Cora and Citeseer networks.

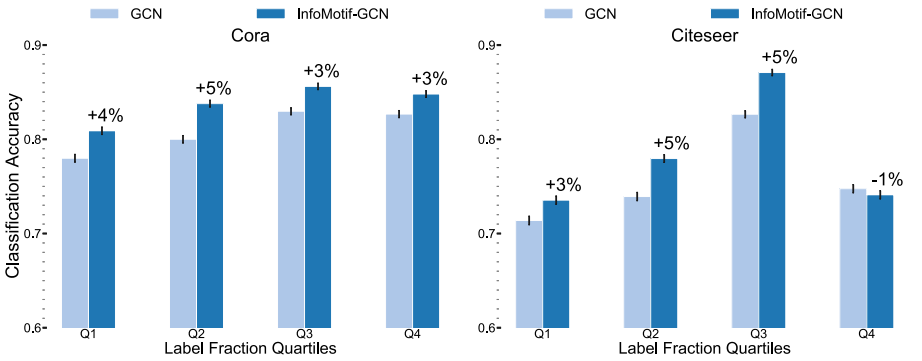


Fig. 7 Classification accuracy over *label fraction* quartiles. (Q1: smaller label fraction). InfoMotif has larger gains over GCN in Q1 and Q2 (nodes that exhibit label sparsity)

Node degree

We divide the set of test nodes into bins based on four degree ranges. Figure 6 depicts the variation in classification accuracy for GCN and InfoMotif-GCN across degree segments, on Cora and Citeseer datasets.

InfoMotif has consistent performance improvements over GCN across all degree segments, with notably higher gains for low-to-medium degree nodes (quartiles Q1 and Q2). Learning structural roles through self-supervised motif regularization is beneficial for nodes with limited local structural information.

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 7 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 (Sect. 4.1), *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 Fig. 8.

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).

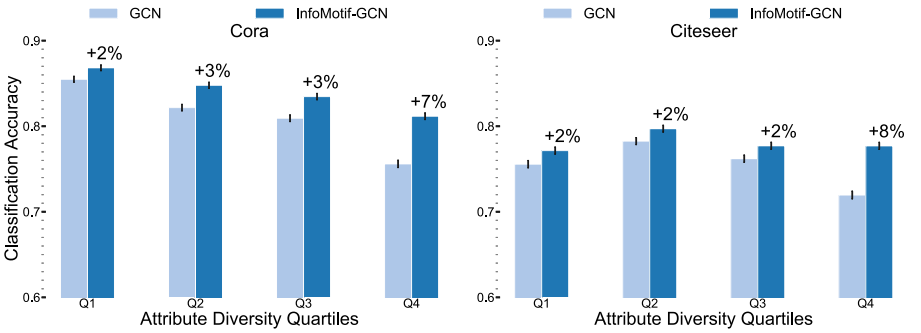


Fig. 8 Classification accuracy across attribute diversity quartiles. (Q4: high attribute diversity). InfoMotif has stronger gains in Q3 and Q4 (nodes with diverse attributed neighborhoods)

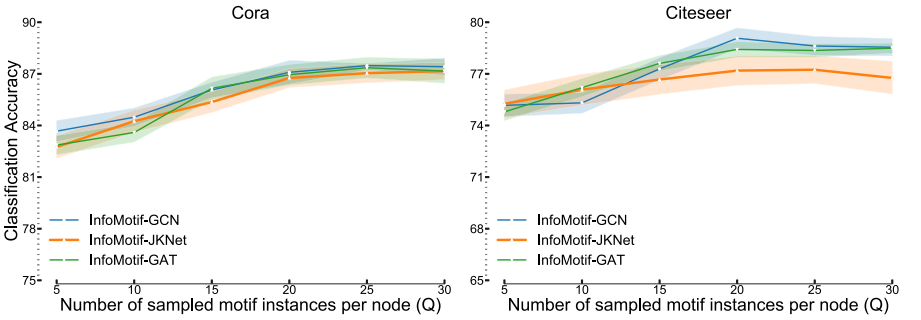


Fig. 9 Classification accuracy increases slowly with the number of sampled motif instances and stabilizes around 15 to 20. Variance bands indicate 95% confidence intervals over 10 runs

6.8 Efficiency and sensitivity analysis (RQ_5)

We now analyze the sensitivity of InfoMotif to hyper-parameters and empirically quantify the cost of motif-based regularization on different choices of base GNN models.

Parameter sensitivity

We examine the effect of hyper-parameter Q that controls the number of motif instances sampled per node to train our motif-based discriminators (Eq. 4). Figure 9 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 stabilizes 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

Efficiency analysis

We empirically evaluate the added complexity of InfoMotif on two base GNN models, GCN, and GAT. We report the model training time per epoch (forward pass, loss computation, backward pass) on synthetically generated Barabasi–Albert networks [1] with 5000 nodes and increasing link density (Fig. 10).

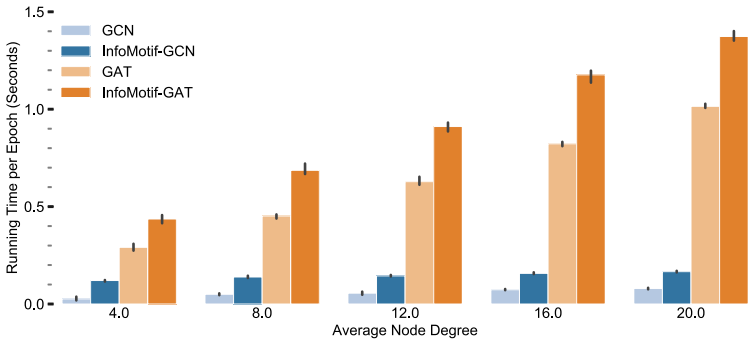


Fig. 10 Runtime comparison of InfoMotif variants with its base GNNs. InfoMotif has minimal computational overheads; notice the nearly constant runtime gap with increasing degree

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 (Fig. 10). Furthermore, our GCN variant InfoMotif-GCN is significantly more efficient than GAT.

7 Discussion

Our framework is designed to address the limitations of *oversmoothing* and *localization* in prior message-passing GNNs.

InfoMotif is orthogonal to advances in GNN architectures that improve the *structural distinguishability* of node representations through carefully designed neighborhood aggregators. In contrast, we enhance the structural resolution of node representations by regularizing base GNNs through self-supervised learning objectives designed to capture connectivity in higher-order motif structures. By training contrastive discriminators to discover attribute correlations among motif instances across the entire graph, our approach learns generalizable *global* roles in addition to modeling local connectivity patterns. The enhanced quality of our learned node representations is further evidenced by our superior empirical performance for nodes with diverse attributed neighborhoods (Sect. 6.7).

InfoMotif addresses the challenge of *localization* by regularizing base GNNs to learn attributed structural roles through self-supervised training objectives. Specifically, our approach statistically relates distant nodes in the graph with covarying attributed structures, to effectively overcome label sparsity in local neighborhoods (Sect. 6.7). Instead of adopting deeper GNNs that directly expand neighborhood aggregation ranges, we demonstrate the effectiveness of regularizing shallow base GNNs to learn attributed structural roles. Compared to deeper GNNs that scale poorly with neighborhood sizes, our regularization strategy enables efficient model inference.

In our work, we choose network motif structures as the central basis to formulate structural roles. In contrast to alternative approaches to quantify structural similarity based on coarse properties like degree sequences [36] or rigid notions of structural equivalence [41], network motifs are fundamental higher-order connectivity structures that enable flexible generalization to complex heterogeneous graphs with rich semantics. We empirically demonstrate the utilities of untyped motifs in homogeneous graphs and typed motifs in heterogeneous graphs.

Our key modeling hypothesis is the importance of attribute covariance in local structures toward the learning application (*e.g.*, classification in social networks). Our substantial gains on two diverse classes of datasets indicate broad applicability for InfoMotif across graphs with varied structural characteristics. However, the performance gains may diminish in application scenarios (*e.g.*, learning in regular mesh graphs) where modeling such covariance is not beneficial or even necessary.

8 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 covarying attributes, independent of network proximity. InfoMotif maximizes motif-based mutual information and dynamically prioritizes the significance of different motifs. Our experiments on nine real-world datasets spanning homogeneous and heterogeneous networks show substantial consistent gains for InfoMotif over state-of-the-art methods.

Acknowledgements We thank anonymous reviewers for their very useful comments and suggestions. Part of this work was done, while Li Shen and Ling Cheng were doing research in Griffith University. The work was supported by Australian Research Council (ARC) Large Grant A849602031.

References

1. Albert R, Barabási AL (2001) Statistical mechanics of complex networks. CoRR, cond-mat/0106096
2. Albert R, Albert-László B (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47
3. Bachman P, Hjelm RD, Buchwalter W (2019) Learning representations by maximizing mutual information across views. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in neural information processing systems 32: annual conference on neural information processing systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pp 15509–15519
4. Belghazi MI, Baratin A, Rajeswar S, Ozair S, Bengio Y, Hjelm RD, Courville AC (2018) Mutual information neural estimation. In: Dy JC, Krause A (eds) Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, volume 80 of Proceedings of Machine Learning Research. PMLR, pp 530–539
5. Beyer L, Zhai X, Oliver A, Kolesnikov A (2019) S4L: self-supervised semi-supervised learning. In: 2019 IEEE/CVF international conference on computer vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019, IEEE, pp 1476–1485
6. Bruna J, Zaremba W, Szlam A, Yann L (2014) Spectral networks and locally connected networks on graphs. In: Bengio Y, LeCun Y (eds) 2nd international conference on learning representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings
7. Cui Y, Jia M, Lin T-Y, Song Y, Belongie SJ (2019) Class-balanced loss based on effective number of samples. In: IEEE conference on computer vision and pattern recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019. Computer Vision Foundation/IEEE, pp 9268–9277
8. Dareddy MR, Das M, Yang H (2019) motif2vec: Motif aware node representation learning for heterogeneous networks. In: Baru C, Huan J, Khan L, Hu X, Ak R, Tian Y, Barga RS, Zaniolo C, Lee K, Ye YF (eds) 2019 IEEE international conference on big data (IEEE BigData), Los Angeles, CA, USA, December 9–12, 2019. IEEE, pp 1052–1059
9. Dauphin YN, Fan A, Auli M, Grangier D (2017) Language modeling with gated convolutional networks. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, volume 70 of Proceedings of Machine Learning Research. PMLR, pp 933–941

10. Dong Y, Chawla NV, Swami A (2017) metapath2vec: scalable representation learning for heterogeneous networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, Halifax, NS, Canada, August 13–17, 2017. ACM, pp 135–144
11. Donnat C, Zitnik M, Hallac D, Leskovec J (2018) Learning structural node embeddings via diffusion wavelets. In: Guo Y, Farooq F (eds) Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, KDD 2018, London, UK, August 19–23, 2018. ACM, pp 1320–1329
12. Fu T-Y, Lee W-C, Lei Z (2017) Hin2vec: explore meta-paths in heterogeneous information networks for representation learning. In: Lim F-P, Winslett M, Sanderson M, Fu AW-C, Sun J, Shane Culpepper J, Lo E, Ho JC, Donato D, Agrawal R, Zheng Y, Castillo C, Sun A, Tseng VC, Li C (eds) Proceedings of the 2017 ACM on conference on information and knowledge management, CIKM 2017, Singapore, November 06–10, 2017. ACM, pp 1797–1806
13. Fu X, Zhang J, Meng Z, King I (2020) MAGNN: metapath aggregated graph neural network for heterogeneous graph embedding. In: Huang Y, King I, Liu T-Y , van Steen M (eds) WWW '20: the web conference 2020, Taipei, Taiwan, April 20–24, 2020. ACM / IW3C2, pp 2331–2341
14. Grover A, Leskovec J (2016) node2vec: Scalable feature learning for networks. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R (eds) Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, USA, August 13–17, 2016. ACM, pp 855–864
15. Hamilton WL, Ying Z, Leskovec J (2017) Inductive representation learning on large graphs. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp 1024–1034
16. Harper FM, Konstan JA (2016) The movielens datasets: history and context. ACM Trans Interact Intell Syst 5(4):19:1-19:19
17. Henderson K, Gallagher B, Eliassi-Rad T, Tong H, Basu S, Akoglu L, Koutra D, Faloutsos C, Li L (2012) Rolx: structural role extraction & mining in large graphs. In: Yang Q, Agarwal D, Pei J (eds) The 18th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '12, Beijing, China, August 12–16, 2012. ACM, pp 1231–1239
18. Hjelm RD, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A, Bengio Y (2019) Learning deep representations by mutual information estimation and maximization. In: 7th international conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net
19. Hu Z, Dong Y, Wang K, Sun Y (2020) Heterogeneous graph transformer. In: Huang Y, King I, Liu T-Y, van Steen M (eds) WWW '20: the web conference 2020, Taipei, Taiwan, April 20–24, 2020. ACM/IW3C2, pp 2704–2710
20. Jin W, Derr T, Liu H, Wang Y, Wang S, Liu Z, Tang J (2020) Self-supervised learning on graphs: deep insights and new direction. CoRR, arXiv:2006.10141
21. Jin W, Derr T, Wang Y, Ma Y, Liu Z, Tang J (2021) Node similarity preserving graph convolutional networks. In: Lewin-Eytan L, Carmel D, Yom-Tov E, Agichtein E, Gabrilovich E (eds) WSDM '21, The fourteenth ACM international conference on web search and data mining, virtual event, Israel, March 8–12, 2021. ACM, pp 148–156
22. Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. OpenReview.net
23. Krishnan A, Cheruvu H, Cheng T, Sundaram H (2019) A modular adversarial approach to social recommendation. In: Zhu W, Tao D, Cheng X, Cui P, Rundensteiner EA, Carmel D, He Q, Yu JX (eds) Proceedings of the 28th ACM international conference on information and knowledge management, CIKM 2019, Beijing, China, November 3–7, 2019. ACM, pp 1753–1762
24. Lee JB, Ryan RA, Kong X, Kim S, Koh E, Rao A (2019) Graph convolutional networks with motif-based attention. In: Zhu W, Tao D, Cheng X, Cui P, Rundensteiner EA, Carmel D, He Q, Yu JX (eds) Proceedings of the 28th ACM international conference on information and knowledge management, CIKM 2019, Beijing, China, November 3–7, 2019. ACM, pp 499–508
25. Li Q, Han Z, Wu X-M (2018) Deeper insights into graph convolutional networks for semi-supervised learning. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018. AAAI Press, pp 3538–3545
26. Liu Y, Pan S, Jin M, Zhou C, Xia F, Yu PS (2021) Graph self-supervised learning: asurvey. CoRR, arXiv:2103.00111

27. McCallum A, Nigam K, Rennie J, Seymore K (2000) Automating the construction of internet portals with machine learning. Inf Retr 3(2):127–163
28. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827
29. Narang K, Yang C, Krishnan A, Wang J, Sundaram H, Sutter C (2019) An induced multi-relational framework for answer selection in community question answer platforms. CoRR, arXiv:1911.06957
30. Paranjape A, Benson AR, Leskovec J (2017) Motifs in temporal networks. In: de Rijke M, Shokouhi M, Tomkins A, Zhang M (eds) Proceedings of the Tenth ACM international conference on web search and data mining, WSDM 2017, Cambridge, United Kingdom, February 6–10, 2017. ACM, pp 601–610
31. Peel L, Delvenne J-C, Lambiotte R (2017) Multiscale mixing patterns in networks. CoRR, arXiv:1708.01236
32. Peng Z, Dong Y, Luo M, Wu X-M, Zheng Q (2020) Self-supervised graph representation learning via global context prediction. CoRR, arXiv:2003.01604
33. Peng Z, Huang W, Luo M, Zheng Q, Rong Y, Xu T, Huang J (2020) Graph representation learning via graphical mutual information maximization. In: Huang Y, King I, Liu T-Y, van Steen M (eds) WWW '20: The Web conference 2020, Taipei, Taiwan, April 20–24, 2020. ACM/IW3C2, pp 259–270
34. Qiu J, Chen Q, Dong Y, Zhang J, Yang H, Ding M, Wang K, Tang J (2020) GCC: graph contrastive coding for graph neural network pre-training. In: Gupta R, Liu Y, Tang J, Aditya Prakash B (eds) KDD '20: The 26th ACM SIGKDD conference on knowledge discovery and data mining, virtual event, CA, USA, August 23–27, 2020. ACM, pp 1150–1160
35. Ren M, Zeng W, Yang B, Urtasun R (2018) Learning to reweight examples for robust deep learning. In: Dy JC, Krause A (eds) Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, volume 80 of Proceedings of Machine Learning Research. PMLR, pp 4331–4340
36. Ribeiro LFR, Saverese PHP, Figueiredo DR (2017) *struc2vec*: learning node representations from structural identity. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, Halifax, NS, Canada, August 13–17, 2017. ACM, pp 385–394
37. Ribeiro P, Paredes P, Silva MEP, Aparício D, Fernando SMA (2021) A survey on subgraph counting: concepts, algorithms, and applications to network motifs and graphlets. ACM Comput Surv 54(2):28:1-28:36
38. Rossi RA, Ahmed NK (2015) Role discovery in networks. IEEE Trans Knowl Data Eng 27(4):1112–1131
39. Rossi RA, Ahmed NK, Carranza AG, Arbour D, Rao A, Kim S, Koh E (2019) Heterogeneous network motifs. CoRR, arXiv:1901.10026
40. Rossi RA, Ahmed NK, Koh E, Kim S, Rao A, Abbasi-Yadkori Y (2020) A structural graph representation learning framework. In: Caverlee J, Hu XB, Lalmas M, Wang W (eds) WSDM '20: The Thirteenth ACM international conference on web search and data mining, Houston, TX, USA, February 3–7, 2020. ACM, pp 483–491
41. Rossi RA, Jin D, Kim S, Ahmed NK, Koutra D, Boaz Lee J (2019) From community to role-based graph embeddings. CoRR, arXiv:1908.08572
42. Rossi RA, Rong Z, Ahmed NK (2019) Estimation of graphlet counts in massive networks. IEEE Trans Neural Netw Learn Syst 30(1):44–57
43. Sankar A (2022) Sparsity-aware neural user behavior modeling in online interaction platforms. Preprint arXiv:2202.13491
44. Sankar A, Liu Y, Yu J, Shah N (2021) Graph neural networks for friend ranking in large-scale social platforms. In: Leskovec J, Grobelnik M, Najork M, Tang J, Zia L (eds) WWW '21: The Web Conference 2021, Virtual Event/Ljubljana, Slovenia, April 19–23, 2021. ACM/IW3C2, pp 2535–2546
45. Sankar A, Wang J, Krishnan A, Sundaram H (2021) Protocf: Prototypical collaborative filtering for few-shot recommendation. In: Jesús Corona Pampín H, Larson MA, Willemsen MC, Konstan JA, McAuley JJ, Garcia-Gathright J, Huurnink B, Oldridge E (eds) *RecSys '21: Fifteenth ACM conference on recommender systems, Amsterdam, The Netherlands, 27 September 2021–1 October 2021*. ACM, pp 166–175
46. Sankar A, Wu Y, Gou L, Zhang W, Yang H (2020) Dysat: Deep neural representation learning on dynamic graphs via self-attention networks. In: Caverlee J, Hu XB, Lalmas M, Wang W (eds) WSDM '20: the thirteenth ACM international conference on web search and data mining, Houston, TX, USA, February 3–7, 2020. ACM, pp 519–527
47. Sankar A, Wu Y, Wu Y, Zhang W, Yang H, Sundaram H (2020) Groupim: a mutual information maximization framework for neural group recommendation. In: Huang J, Chang Y, Cheng X, Kamps J, Murdock V, Wen J-R, Liu Y (eds) Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25–30, 2020. ACM, pp 1279–1288

48. Sankar A, Zhang X, Chen-Chuan Chang K (2017) Motif-based convolutional neural network on graphs. CoRR, arXiv:1711.05697
49. Sankar A, Zhang X, Chen-Chuan Chang K (2019) Meta-gnn: metagraph neural network for semi-supervised learning in attributed heterogeneous information networks. In: Spezzano F, Chen W, Xiao X (eds) ASONAM '19: international conference on advances in social networks analysis and mining, Vancouver, British Columbia, Canada, 27–30 August, 2019. ACM, pp 137–144
50. Sankar A, Zhang X, Krishnan A, Han J (2020) Inf-vae: a variational autoencoder framework to integrate homophily and influence in diffusion prediction. In: Caverlee J, Hu XB, Lalmas M, Wang W (eds) WSDM '20: the thirteenth ACM international conference on web search and data mining, Houston, TX, USA, February 3–7, 2020. ACM, pp 510–518
51. Sen P, Namata G, Bilgic M, Getoor L, Galligher B, Eliassi-Rad T (2008) Collective classification in network data. AI Mag 29(3):93–106
52. Shchur O, Mumme M, Bojchevski A, Günnemann S (2018) Pitfalls of graph neural network evaluation. CoRR, arXiv:1811.05868
53. Shi Y, Gui H, Zhu Q, Lance KM, Han J (2018) Aspem: Embedding learning by aspects in heterogeneous information networks. In: Ester M, Pedreschi D (eds) Proceedings of the 2018 SIAM international conference on data mining, SDM 2018, May 3–5, 2018, San Diego Marriott Mission Valley, San Diego, CA, USA. SIAM, pp 144–152
54. Sun Y, Han J, Yan X, Yu PS, Wu T (2011) Pathsim: meta path-based top-k similarity search in heterogeneous information networks. Proc VLDB Endow 4(11):992–1003
55. Tu K, Cui P, Wang X, Philip YS, Zhu W (2018) Deep recursive network embedding with regular equivalence. In: Guo Y, Farooq F (eds) Proceeding of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, KDD 2018, London, UK, August 19–23, 2018
56. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Aidan GN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp 5998–6008
57. Velickovic P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y (2018) Graph attention networks. In: 6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings. OpenReview.net
58. Velickovic P, Fedus W, Hamilton WL, Liò P, Bengio Y, Hjelm DR (2019) Deep graph infomax. In: 7th international conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net
59. Vinyals O, Bengio S, Kudlur M (2016) Order matters: sequence to sequence for sets. In: Bengio Y, LeCun Y (eds) 4th international conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings
60. Wang X, Ji H, Shi C, Wang B, Ye Y, Cui P, Philip YS (2019) Heterogeneous graph attention network. In: Liu L, White RW, Mantrach A, Silvestri F, McAuley JJ, Baeza-Yates R, Zia L (eds) The world wide web conference, WWW 2019, San Francisco, CA, USA, May 13–17, 2019. ACM, pp 2022–2032
61. Wu J, He J, Xu J (2019) Demo-net: segree-specific graph neural networks for node and graph classification. In: Teredesai A, Kumar V, Li Y, Rosales R, Terzi E, Karypis G (eds) Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, KDD 2019, Anchorage, AK, USA, August 4–8, 2019. ACM, pp 406–415
62. Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY (2021) A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 32(1):4–24
63. Xie Y, Xu Z, Wang Z, Ji S (2021) Self-supervised learning of graph neural networks: a unified review. CoRR, arXiv:2102.10757
64. Xu K, Li C, Tian Y, Sonobe T, Kawarabayashi K-I, Jegelka S (2018) Representation learning on graphs with jumping knowledge networks. In: Dy JC, Krause A (eds) Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, volume 80 of Proceedings of Machine Learning Research. PMLR, pp 5449–5458
65. You J, Ying R, Leskovec J (2019) Position-aware graph neural networks. In: Chaudhuri K, Salakhutdinov R(eds) Proceedings of the 36th international conference on machine learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research. PMLR, pp 7134–7143
66. You Y, Chen T, Sui Y, Chen T, Wang Z, Shen Y (2020) Graph contrastive learning with augmentations. In: Larochelle H, Ranzato M, Hadsell R, Balcan M-F, Lin H-T (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual

67. Zhang C, Song D, Huang C, Swami A, Nitesh CV (2019) Heterogeneous graph neural network. In: Teredesai A, Kumar V, Li Y, Rosales R, Terzi E, Karypis G (eds) Proceedings of the 25th ACM SIGKDD International conference on knowledge discovery & data mining, KDD 2019, Anchorage, AK, USA, August 4–8, 2019. ACM. pp 793–803
68. Zhang D, Yin J, Zhu X, Zhang C (2018) Metagraph2vec: Complex semantic path augmented heterogeneous network embedding. In: Phung DQ, Tseng VS, Webb GI, Ho B, Ganji M, Rashidi L (eds) Advances in knowledge discovery and data mining—22nd pacific-asia conference, PAKDD 2018, Melbourne, VIC, Australia, June 3–6, 2018, Proceedings, Part II, volume 10938 of Lecture Notes in Computer Science. Springer, pp 196–208
69. Zhang Y, Xiong Y, Kong X, Li S, Mi J, Zhu Y (2018) Deep collective classification in heterogeneous information networks. In: Champin P-A, Gandon F, Lalmas M, Ipeirotis PG (eds) Proceedings of the 2018 world wide web conference on world wide web, WWW 2018, Lyon, France, April 23–27, 2018. ACM, pp 399–408
70. Zhou D, Bousquet O, Navin TL, Weston J, Schölkopf B (2003) Learning with local and global consistency. In: Thrun S, Saul LK, Schölkopf B (eds) Advances in neural information processing systems 16 [neural information processing systems, NIPS 2003, December 8–13, 2003, Vancouver and Whistler, British Columbia, Canada]. MIT Press, pp 321–328
71. Zhou X, Belkin M (2014) Semi-supervised learning. In: Academic press library in signal processing. Elsevier, vol 1, pp 1239–1269
72. Zhuang C, Ma Q (2018) Dual graph convolutional networks for graph-based semi-supervised classification. In: Champin P-A, Gandon F, Lalmas M, Ipeirotis PG (eds) Proceedings of the 2018 world wide web conference on world wide web, WWW 2018, Lyon, France, April 23–27, 2018. ACM, pp 499–508

Aravind Sankar is a research scientist in the Core Data Science team at Meta Research. He received his Bachelors and Masters in Computer Science at the University of Illinois at Urbana-Champaign in 2020 and 2022 respectively. His PhD dissertation work focused on robust user behavior modeling in large-scale social networking and online interaction platforms. His research interests span data mining, applied machine learning, and computational social science domains, with a focus on recommender systems and graph mining. He has designed graph neural networks and neural collaborative filtering models for graph representation learning, personalized recommendation, and social interaction modeling applications.

Junting Wang is an incoming PhD student at the University of Illinois, Urbana-Champaign. He received his Bachelors and Masters in Computer Science at the University of Illinois at Urbana-Champaign in 2020 and 2022 respectively. He has designed personalized recommendation and graph representation learning models across several applications. His current research directions include solving fundamental research problems, *e.g.*, data sparsity, bias and fairness, scalability, in the field of data mining, recommender systems, graph mining, and information retrieval.

Adit Krishnan is an Applied Scientist at Microsoft focused on multimodal content retrieval, analysis, and recommendation. He received his PhD in Computer Science at the University of Illinois at Urbana-Champaign in 2021. His research primarily develops robust neural network architectures, models, and training methodologies to tackle data challenges associated with machine learning in large-scale search, recommendation, and retrieval applications. His research work has been recognized with an Amazon Research Award in 2019. He has also served on the program committees of multiple top-tier academic conferences in the data mining and information retrieval domains.

Hari Sundaram is a professor in the Computer Science Department at the University of Illinois at Urbana-Champaign with affiliate appointments in the Charles H. Sandage Department of Advertising, the Institute for Communication Research and the Center for Social & Behavioral Science. Sundaram received his PhD in Electrical Engineering at Columbia University (2002). He received the Eliahu Jury award for best dissertation (2002), IBM faculty awards (2007, 2008), and several best-paper awards and best-paper runner-up honors from IEEE and ACM conferences. He was elected as ACM distinguished scientist in 2019 and IEEE senior member in 2019. Prof. Sundaram's research spans applied machine learning, network science, and human–computer interaction. He develops algorithms and builds systems that help individuals to understand and to act.