Heterogeneous Graph Information Bottleneck

Liang Yang [1,2,3], **Fan Wu** [1], **Zichen Zheng** [1], **Bingxin Niu** [1,3], **Junhua Gu** [1,3], **Chuan Wang** [2], **Xiaochun Cao** [2] and **Yuanfang Guo** [4*]

[1]School of Artificial Intelligence, Hebei University of Technology, Tianjin, China
[2]State Key Laboratory of Information Security, Institute of Information Engineering, CAS, Beijing, China
[3]Hebei Province Key Laboratory of Big Data Calculation, Hebei University of Technology, China
[4]Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Computer Science and Engineering, Beihang University, Beijing, China
yangliang@vip.qq.com, andyguo@buaa.edu.cn

Abstract

Most attempts on extending Graph Neural Networks (GNNs) to Heterogeneous Information Networks (HINs) implicitly take the direct assumption that the multiple homogeneous attributed networks induced by different meta-paths are *complementary*. The doubts about the hypothesis of complementary motivate an alternative assumption of *consensus*. That is, the aggregated node attributes shared by multiple homogeneous attributed networks are essential for node representations, while the specific ones in each homogeneous attributed network should be discarded. In this paper, a novel Heterogeneous Graph Information Bottleneck (HGIB) is proposed to implement the consensus hypothesis in an unsupervised manner. To this end, information bottleneck (IB) is extended to unsupervised representation learning by leveraging self-supervision strategy. Specifically, HGIB simultaneously maximizes the mutual information between one homogeneous network and the representation learned from another homogeneous network, while minimizes the mutual information between the specific information contained in one homogeneous network and the representation learned from this homogeneous network. Model analysis reveals that the two extreme cases of HGIB correspond to the supervised heterogeneous GNN and the infomax on homogeneous graph, respectively. Extensive experiments on real datasets demonstrate that the consensus-based unsupervised HGIB significantly outperforms most semi-supervised SOTA methods based on complementary assumption.

1 Introduction

Heterogeneous Information Networks (HINs) possess the advantage of modeling rich relations in real work compared to homogeneous networks, which have been well studied by the researchers from mathematics, physics and computer science [Shi *et al.*, 2017; Wang *et al.*, 2020]. Thus, by effectively exploiting these multiple relations via meta-paths, HINs provide more clues for accurate network analysis, e.g. network embedding [Dong *et al.*, 2017], and have been successfully applied to recommendation system [Shi *et al.*, 2019], natural language processing [Hu *et al.*, 2019] and knowledge graph.

Graph neural networks (GNNs) [Wu *et al.*, 2021], especially graph convolutional neural networks (GCNNs) [Kipf and Welling, 2017; Bruna *et al.*, 2014], have became a powerful tool for homogeneous attributed network embedding. And, their success can be attributed to the Laplacian smoothing [Li *et al.*, 2018] from spatial perspective or the low-pass filtering [Wu *et al.*, 2019] from spectral perspective.

Recent attempts extend GNNs to heterogeneous information networks [Wang *et al.*, 2019; Fu *et al.*, 2020; Yun *et al.*, 2019; Hu *et al.*, 2020]. Most of them follow the pipeline of transforming a heterogeneous attributed network with multiple relations into multiple attributed homogeneous networks via meta-paths and combining the embedding results of multiple homogeneous attributed networks obtained from GNNs. And, the supervision information is utilized to learn how to map from node feature to label and how to combine the multiple embedding results [Wang *et al.*, 2019; Yun *et al.*, 2019]. These semi-supervised methods implicitly take the direct assumption that the multiple homogeneous attributed networks induced by different meta-paths are **complementary**. That is, the information contained in each homogeneous attributed network is insufficient to represent nodes, thus, multiple homogeneous attributed networks are necessary to complete the information.

Here, the direct assumption of complementarity is investigated. The doubts about this hypothesis stem from both the characteristic of the homogeneous attributed networks and the nature of the adopted GNNs. First, the homogeneous attributed networks induced by meta-paths are not independent. In fact, they share the common node attributes (feature) and possess different network topologies. Second, the essence of GNNs, which are applied to each homogeneous attributed network, is the attributes smoothing according to the topology, i.e., discarding noises. Based on these two characteristics, the same attributes are smoothed according to the different topologies of multiple homogeneous networks. Thus, the smoothed node attributes in each homogeneous attributed network may not be significantly different.

*Corresponding author.

Therefore, contrary to hypothesis of complementarity, another alternative assumption may be the **consensus**, where *the aggregated node attributes shared by multiple homogeneous attributed networks are essential for node representations*. In other words, to seek robust node representation, the aggregated node attributes, which are specific in each homogeneous attributed network, should be discarded. This assumption shares the common philosophy with consensus clustering. Note that this alternative assumption reduces the requirement for labels, thus is more suitable for unsupervised tasks.

In this paper, a novel Heterogeneous Graph Information Bottleneck (HGIB) is proposed to implement the consensus hypothesis in an unsupervised manner. To this end, information bottleneck (IB), which has been widely used in supervised tasks, is extended to unsupervised representation learning by leveraging self-supervision strategy. That is, each induced homogeneous attributed network is utilized as the self-supervision information for the representation learning task on other induced homogeneous attributed networks. Specifically, HGIB simultaneously maximizes the mutual information between the representation learned from one homogeneous network and another homogeneous network, and minimizes the mutual information between the specific information contained in one homogeneous network and the representation learned from this homogeneous network. The model analysis reveals that HGIB degrades to the supervised heterogeneous GNN or the infomax on homogeneous graph, respectively, if the two adopted meta-paths are extremely similar or dissimilar.

The main contributions are summarized as follows.

- We investigate the widely-adopted complementary assumption in designing GNNs for HINs, and present an alternative one, i.e., consensus hypothesis, which is more suitable for unsupervised tasks.
- We propose a well-behavior Heterogeneous Graph Information Bottleneck (HGIB) by leveraging self-supervised learning strategy, which facilitates the adoption of information bottleneck for unsupervised tasks.
- We reveal that the two extreme cases of HGIB correspond to the supervised heterogeneous GNN and the infomax on homogeneous graph, respectively.
- Extensive experiments demonstrate that the consensus-based unsupervised HGIB significantly outperforms most semi-supervised SOTA methods based on complementary assumption.

2 Preliminaries

2.1 Heterogeneous Information Network

A heterogeneous information network (HIN) [Sun and Han, 2012], denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi, \varphi)$, consists of a node set $\mathcal{V}$ and a link set $\mathcal{E}$ associating with a node type mapping function $\phi : \mathcal{V} \mapsto \mathcal{T}$ and link type mapping function $\varphi : \mathcal{E} \mapsto \mathcal{R}$, respectively. In the network, each object $v \in \mathcal{V}$ belongs to one specific object type $\phi(v) \in \mathcal{T}$ and each link $e \in \mathcal{E}$ belongs to a specific relation $\varphi(e) \in \mathcal{R}$, where $|\mathcal{T}| + |\mathcal{R}| > 2$. A meta-path $\mathcal{P}$ of length l is denoted in the form of $T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} ... \xrightarrow{R_l} T_{l+1}$, which defines a composite relation $R = R_1 \circ R_2 \circ ... \circ R_l$ between types T_1 and T_{l+1} with $\circ$ standing for the composition operator on relations.

2.2 Information Bottleneck

To investigate the discriminative ability of the representation, the amount of label information that remains accessible after encoding the data, is known as sufficiency [Achille and Soatto, 2018]. A representation $\mathbf{h}$ of data $\mathbf{x}$ is sufficient for the label $\mathbf{y}$ if and only if $I(\mathbf{x}; \mathbf{y}|\mathbf{h}) = 0$. That is, the amount of information regarding the task is unchanged by the encoding procedure, i.e.

$$I(\mathbf{x}; \mathbf{y}) = I(\mathbf{h}; \mathbf{y}). \tag{1}$$

where $I(\cdot;\cdot)$ stands for the mutual information. To make the representation robustness (generalization), Information Bottleneck principle (IB) [Tishby *et al.*, 2000] attempts to discard all information from the input, which is not helpful for a given task. To this end, IB [Alemi *et al.*, 2017] directly minimizes the mutual information between the data $\mathbf{x}$ and its representation $\mathbf{h}$, $I(\mathbf{x}; \mathbf{h})$, while at the same time maximizes the mutual information between $\mathbf{h}$ and the label $\mathbf{y}$, $I(\mathbf{y}; \mathbf{h})$. Its objective function can be formulated as follows

$$R_{IB}(\theta) = I_\theta(\mathbf{y}; \mathbf{h}) - \beta I_\theta(\mathbf{x}; \mathbf{h}). \tag{2}$$

where θ denotes the parameters of the representation encoder $p_\theta(\mathbf{h}|\mathbf{x})$ and β controls the tradeoff. The second term $I(\mathbf{x}; \mathbf{h})$ can be subdivided into two components by using the chain rule of mutual information as

$$I(\mathbf{x}; \mathbf{h}) = I(\mathbf{x}; \mathbf{h}|\mathbf{y}) + I(\mathbf{y}; \mathbf{h}), \tag{3}$$

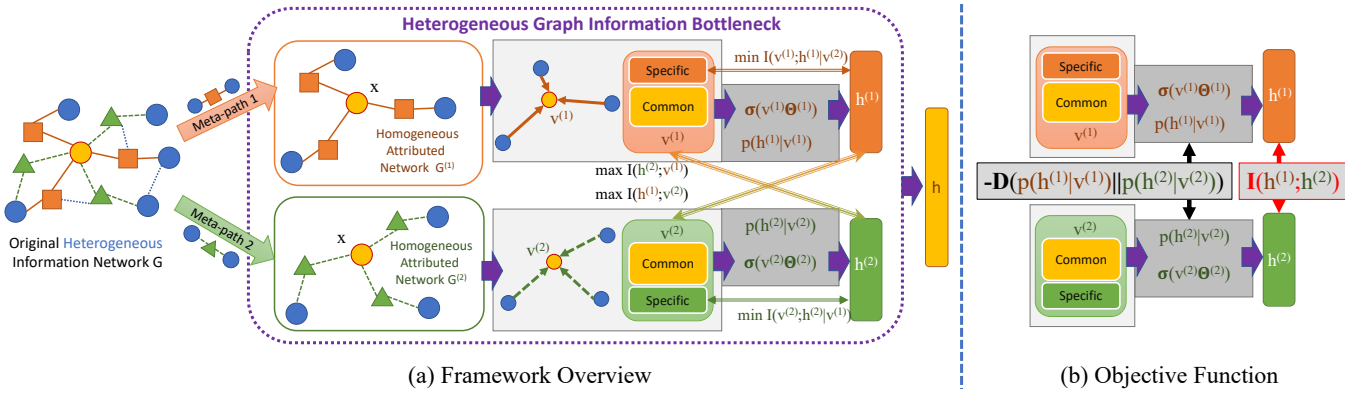
where the second term $I(\mathbf{y}; \mathbf{h})$ is independent of the representation $\mathbf{h}$, since $\mathbf{h}$ is sufficient for $\mathbf{y}$ as shown in Eq. (1). The first term $I(\mathbf{x}; \mathbf{h}|\mathbf{y})$ represents the information in $\mathbf{h}$ that is not predictive of $\mathbf{y}$, i.e. superfluous information. Therefore, minimizing the mutual information $I(\mathbf{x}; \mathbf{h})$ is equivalent to minimizing the superfluous information $I(\mathbf{x}; \mathbf{h}|\mathbf{y})$ [Federici *et al.*, 2020], and the objective of IB in Eq. (4) can be reformulated as

$$R_{IB}(\theta) = I(\mathbf{y}; \mathbf{h}) - \beta I(\mathbf{x}; \mathbf{h}|\mathbf{y}). \tag{4}$$

Note that maximizing the IB can be done directly only in supervised settings, i.e. $\mathbf{y}$ is given.

3 Heterogeneous Graph Information Bottleneck

In this section, Heterogeneous Graph Information Bottleneck (HGIB) is proposed. First, the assumption and the overview are provided by transforming the unsupervised heterogeneous graph neural network as a self-supervised task. Then, the formula of the self-supervised information bottleneck is introduced based on the supervised one in Sec 2.2. Finally, the objective function and the optimization are elaborated.



(a) Framework Overview (b) Objective Function

Figure 1: The illustration of the proposed Heterogeneous Graph Information Bottleneck (HGIB) and its objective function.

3.1 Assumption and Overview

There may exist multiple relations between each pair of nodes, which are induced by different meta-paths, in heterogeneous. Taking ACM as an example, each pair of papers can be connected by the same author or same subject. Contrary to the complementary assumption taken by most existing GNNs for heterogeneous, another alternative assumption, i.e., consensus, is investigated here. Consensus assumption considers that the learned node representations shared by multiple sub-graphs are essential for node representations. In other words, to seek robust node representation, the node representations, which are specific in each sub-graph, should be discarded.

To implement the consensus assumption, the Heterogeneous Graph Information Bottleneck (HGIB) is proposed. Its illustration is shown in Fig. 1. In the heterogeneous graph, circle, square and triangle denote three kinds of nodes, while solid, dashed and dotted lines stand for three kinds of edges. First, the heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi, \varphi)$ is decomposed into two sub-graphs $\mathcal{G}^{(1)} = \left(\bar{\mathcal{V}}, \mathcal{E}^{(1)}\right)$ (upper graph) and $\mathcal{G}^{(2)} = \left(\bar{\mathcal{V}}, \mathcal{E}^{(2)}\right)$ (lower graph) according to the meta-paths "circle-square-circle" and "circle-triangle-circle", respectively, where $\bar{\mathcal{V}}$ represents the set of nodes with the type of circle. The adjacency matrices of these two sub-graphs are denoted as $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$. The attributes of the nodes of circle type are collected in matrix $\mathbf{X}$, and $\mathbf{x}$ is employed to represent the original attributes of one node.

Here, the widely-adopted GCN [Kipf and Welling, 2017] is adopted as the encoders to obtain the node representations on two sub-graphs, as shown in the four gray boxes in Fig. 1, where the light gray boxes and dark gray boxes represent the propagations without learnable parameter and trainable mapping functions, respectively. The formula is as follows

$$\begin{aligned}
\mathbf{H}^{(1)} &= p(\mathbf{H}^{(1)}|\mathbf{V}^{(1)}) = \sigma\left(\mathbf{V}^{(1)}\mathbf{\Theta}^{(1)}\right) \\
&= \sigma\left(\left(\tilde{\mathbf{D}}^{(1)}\right)^{-\frac{1}{2}}\tilde{\mathbf{A}}^{(1)}\left(\tilde{\mathbf{D}}^{(1)}\right)^{-\frac{1}{2}}\mathbf{X}\mathbf{\Theta}^{(1)}\right), \\
\mathbf{H}^{(2)} &= p(\mathbf{H}^{(2)}|\mathbf{V}^{(2)}) = \sigma\left(\mathbf{V}^{(2)}\mathbf{\Theta}^{(2)}\right) \\
&= \sigma\left(\left(\tilde{\mathbf{D}}^{(2)}\right)^{-\frac{1}{2}}\tilde{\mathbf{A}}^{(2)}\left(\tilde{\mathbf{D}}^{(2)}\right)^{-\frac{1}{2}}\mathbf{X}\mathbf{\Theta}^{(2)}\right),
\end{aligned}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ stands for the adjacency matrix with self-loop, $\tilde{\mathbf{D}}$ denotes the degree matrix of $\tilde{\mathbf{A}}$ with the diagonal elements as the degrees of the nodes, $\mathbf{V} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}$ stands for the representations after propagation but without learnable parameters, and $\mathbf{\Theta}$ represents the learnable parameters (The matrices with superscripts $\cdot^{(1)}$ and $\cdot^{(2)}$ correspond to sub-graph $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$, respectively). Besides, $\mathbf{v}$ and $\mathbf{h}$, which are the rows of $\mathbf{V}$ and $\mathbf{H}$, respectively, are used to represent the attributes after propagation and final representation corresponding to one node, respectively. $\sigma(\cdot)$ denotes the nonlinear mapping function, such as ReLU or softmax.

According to the consensus assumption mentioned above, both $\mathbf{v}^{(1)}$ (light orange box) and $\mathbf{v}^{(2)}$ (light green box) share some common and inherent characteristics (yellow part) and possess specific characteristics (dark orange and dark green components), as shown in Fig. 1. Thus, we would like to learn the representation $\mathbf{h}^{(1)}$ (or $\mathbf{h}^{(2)}$) from $\mathbf{v}^{(1)}$ (or $\mathbf{v}^{(2)}$) that discards as much information as possible without losing any label information. In the next subsection, the IB for supervised task provided in Sec. 2.2 will be extended to self-supervised one for discarding as much specific information as possible in the heterogeneous graph information bottleneck.

3.2 Self-supervised Information Bottleneck

In this section, the assumption and overview will be formulated by extending semi-supervised IB to self-supervised one. First, the consensus assumption can be formalized as the redundancy: $\mathbf{v}^{(1)}$ is redundant with respect to $\mathbf{v}^{(2)}$ for $\mathbf{y}$ if and only if $I(\mathbf{y}; \mathbf{v}^{(1)}|\mathbf{v}^{(2)}) = 0$. Whenever $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ are mutually redundant, any representation which contains all the information shared by both is as predictive as their joint observation.

Second, heterogeneous graph information bottleneck (HGIB) is formalized via self-supervised information bottleneck (SSIB). SSIB extends IB by considering the mutual redundancy assumption. Note that HGIB aims at exploring shared information by discarding as much specific information as possible. The IB formula in Eq. (4) can be extend to

$$R_{SSIB}(\theta) = I_\theta(\mathbf{v^{(2)}}; \mathbf{h^{(1)}}) - \beta I_\theta(\mathbf{v^{(1)}}; \mathbf{h^{(1)}}). \quad (5)$$

where the first term $I(\mathbf{v^{(2)}}; \mathbf{h^{(1)}})$ maximizes the shared information between the learned representation $\mathbf{h}^{(1)}$ from sub-

graph $\mathcal{G}^{(1)}$ and $\mathbf{v}^{(2)}$ from sub-graph $\mathcal{G}^{(2)}$, while the second term $I(\mathbf{v}^{(1)};\mathbf{h}^{(1)})$ minimizes the information in $\mathbf{h}^{(1)}$) from data $\mathbf{v}^{(1)}$. Similar to the decomposition in Eq. (3), by employing the redundancy assumption, the second term in SSIB (Eq. (5)) can be decomposed into

$$I(\mathbf{v}^{(1)};\mathbf{h}^{(1)}) = I(\mathbf{v}^{(1)};\mathbf{h}^{(1)}|\mathbf{v}^{(2)}) + I(\mathbf{v}^{(2)};\mathbf{h}^{(1)}). \quad (6)$$

Since $I(\mathbf{v}^{(2)};\mathbf{h}^{(1)})$ needs to be maximal to maximize R_{SSIB} in Eq. (5), to minimize $I(\mathbf{v}^{(1)};\mathbf{h}^{(1)})$, the first term in Eq. (6), i.e., $I(\mathbf{v}^{(1)};\mathbf{h}^{(1)}|\mathbf{v}^{(2)})$, which represents the information in $\mathbf{h}^{(1)}$ from the specific information of $\mathbf{v}^{(1)}$, should be minimized. Thus, the R_{SSIB} in Eq. (5) can be rewritten as

$$R^{(1)}_{SSIB}(\Theta^{(1)}) = I_{\Theta^{(1)}}(\mathbf{v}^{(2)};\mathbf{h}^{(1)}) - \beta_1 I_{\Theta^{(1)}}(\mathbf{v}^{(1)};\mathbf{h}^{(1)}|\mathbf{v}^{(2)}), \quad (7)$$

where $\Theta^{(1)}$ denotes the parameter of the GCN applied on sub-graph $\mathcal{G}^{(1)}$. Note that, in Fig. 1, the first term is illustrated as the oblique line between $\mathbf{v}^{(2)}$ and $\mathbf{h}^{(1)}$, while the second term as the horizontal line between $\mathbf{h}^{(1)}$ and the specific part in $\mathbf{v}^{(1)}$. Similarly, the counterpart of $R^{(2)}_{SSIB}$ can be formulated as

$$R^{(2)}_{SSIB}(\Theta^{(2)}) = I_{\Theta^{(2)}}(\mathbf{v}^{(1)};\mathbf{h}^{(2)}) - \beta_2 I_{\Theta^{(2)}}(\mathbf{v}^{(2)};\mathbf{h}^{(2)}|\mathbf{v}^{(1)}). \quad (8)$$

Therefore, the self-supervised information bottleneck is formulated as jointly maximizing both $R^{(1)}_{SSIB}(\Theta^{(1)})$ and $R^{(2)}_{SSIB}(\Theta^{(2)})$, and the two representation encoders implemented via GCNs with parameters $\mathbf{\Theta}^{(1)}$ and $\mathbf{\Theta}^{(2)}$ can be obtained. Furthermore, the final node representation is obtained as $\mathbf{h} = \mathbf{h}^{(1)} + \mathbf{h}^{(2)}$, where the $+$ stands for the element-wise summation.

3.3 Objective Function and Optimization

Objective Function: As described in pervious subsection, the objective function of HGIB can be obtained by averaging Eqs. (7) and (8) as

$$\begin{aligned} & R_{SSIB}(\Theta^{(1)},\Theta^{(2)}) = \frac{1}{2}\left(R^{(1)}_{SSIB} + R^{(2)}_{SSIB}\right) \\ = \; - \; & \frac{\beta_1 I_{\Theta^{(1)}}(\mathbf{v}^{(1)};\mathbf{h}^{(1)}|\mathbf{v}^{(2)}) + \beta_2 I_{\Theta^{(2)}}(\mathbf{v}^{(2)};\mathbf{h}^{(2)}|\mathbf{v}^{(1)})}{2} \\ + \; & \frac{I_{\Theta^{(1)}}(\mathbf{v}^{(2)};\mathbf{h}^{(1)}) + I_{\Theta^{(2)}}(\mathbf{v}^{(1)};\mathbf{h}^{(2)})}{2}. \end{aligned} \quad (9)$$

On one hand, by considering $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$ on the same domain, the term $I_{\Theta^{(1)}}(\mathbf{v}^{(1)};\mathbf{h}^{(1)}|\mathbf{v}^{(2)})$ can be expressed as

$$\begin{aligned} & I_{\Theta^{(1)}}(\mathbf{v}^{(1)};\mathbf{h}^{(1)}|\mathbf{v}^{(2)}) \\ \leq \; & D_{KL}(p_{\Theta^{(1)}}(\mathbf{h}^{(1)}|\mathbf{v}^{(1)})||p_{\Theta^{(2)}}(\mathbf{h}^{(2)}|\mathbf{v}^{(2)})), \end{aligned} \quad (10)$$

which provide an upper-bound. This bound is tight when $p_{\Theta^{(1)}}(h^{(1)}|v^{(1)})$ coincides with $p_{\Theta^{(2)}}(h^{(2)}|v^{(2)})$, i.e., the two GCNs provide a consistent representation encoder. The derivation is provided in supplementary. Similarly, the counterpart is

$$\begin{aligned} & I_{\Theta^{(2)}}(\mathbf{v}^{(2)};\mathbf{h}^{(2)}|\mathbf{v}^{(1)}) \\ \leq \; & D_{KL}(p_{\Theta^{(2)}}(\mathbf{h}^{(2)}|\mathbf{v}^{(2)})||p_{\Theta^{(1)}}(\mathbf{h}^{(1)}|\mathbf{v}^{(1)})). \end{aligned} \quad (11)$$

On the other hand, $I_{\Theta^{(1)}}(\mathbf{v}^{(2)};\mathbf{h}^{(1)})$ can be reformulated as follows

$$I_{\Theta^{(1)}}(\mathbf{v}^{(2)};\mathbf{h}^{(1)}) \geq I_{\Theta^{(1)}\Theta^{(2)}}(\mathbf{h}^{(1)};\mathbf{h}^{(2)}), \quad (12)$$

where the second equality sign is because $\mathbf{h}^{(2)}$ is the representation of $\mathbf{v}^{(2)}$. This bound is tight when $\mathbf{h}^{(2)}$ is sufficient for $\mathbf{h}^{(1)}$, i.e., $\mathbf{h}^{(2)}$ contains all the information regarding $\mathbf{h}^{(1)}$. The derivation is provided in supplementary. Analogously, it holds that

$$I_{\Theta^{(2)}}(\mathbf{v}^{(1)};\mathbf{h}^{(2)}) \geq I_{\Theta^{(1)}\Theta^{(2)}}(\mathbf{h}^{(1)};\mathbf{h}^{(2)}). \quad (13)$$

Therefore, the objective function in Eq. (14) can be lower-bounded as

$$\begin{aligned} & R_{SSIB}(\Theta^{(1)},\Theta^{(2)}) \\ \geq \; - \; & \gamma D_{SKL}(p_{\Theta^{(1)}}(\mathbf{h}^{(1)}|\mathbf{v}^{(1)})||p_{\Theta^{(2)}}(\mathbf{h}^{(2)}|\mathbf{v}^{(2)})) \\ + \; & I_{\Theta^{(1)}\Theta^{(2)}}(\mathbf{h}^{(1)};\mathbf{h}^{(2)}), \end{aligned} \quad (14)$$

where $D_{SKL}(p_{\Theta^{(1)}}(\mathbf{h}^{(1)}|\mathbf{v}^{(1)})||p_{\Theta^{(2)}}(\mathbf{h}^{(2)}|\mathbf{v}^{(2)}))$ stands for the symmetric KL divergence between $p_{\Theta^{(1)}}(\mathbf{h}^{(1)}|\mathbf{v}^{(1)})$ and $p_{\Theta^{(2)}}(\mathbf{h}^{(2)}|\mathbf{v}^{(2)})$. γ is employed to control the trade-off between sufficiency and robustness of the node representation. For clarity, this objective function is illustrated in Fig.1(b).

Optimization: First, the symmetric KL divergence $D_{SKL}(p_{\Theta^{(1)}}(\mathbf{h}^{(1)}|\mathbf{v}^{(1)})||p_{\Theta^{(2)}}(\mathbf{h}^{(2)}|\mathbf{v}^{(2)}))$ can be directly computed by setting both $p_{\Theta^{(1)}}(\mathbf{h}^{(1)}|\mathbf{v}^{(1)})$ and $p_{\Theta^{(2)}}(\mathbf{h}^{(2)}|\mathbf{v}^{(2)})$ as Gaussian distributions whose mean and diagonal covariance matrix are obtained by feeding $\mathbf{v}$ ($\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ are for two probability distributions, respectively) into a fully-connected layer. Second, the mutual information $I_{\Theta^{(1)}\Theta^{(2)}}(\mathbf{h}^{(1)};\mathbf{h}^{(2)})$ between the node representations $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$ can be maximized by using any sample-based differentiable mutual information lower bound, such as MINE [Belghazi *et al.*, 2018] and inforNCE [Oord *et al.*, 2018]. Here, MINE is employed to estimate the mutual information as

$$\begin{aligned} & \hat{I}_{\Theta^{(1)}\Theta^{(2)}}(\mathbf{h}^{(1)};\mathbf{h}^{(2)}) \\ = \; & \sup_{\psi} \mathbb{E}_{p(\mathbf{h}^{(1)},\mathbf{h}^{(2)})}[T_\psi] - \log\left(\mathbb{E}_{p(\mathbf{h}^{(1)})p(\mathbf{h}^{(2)})}[e^{T_\psi}]\right) \end{aligned} \quad (15)$$

where $T_\psi = T_\psi(\mathbf{h}^{(1)},\mathbf{h}^{(2)})$ is a function parametrized by a deep neural network with parameters ψ, and a multi-layer perceptron (MLP) is employed in this paper. $p(\mathbf{h}^{(1)},\mathbf{h}^{(2)})$ is joint sampler, while $p(\mathbf{h}^{(1)})$ and $p(\mathbf{h}^{(2)})$ are marginal ones.

3.4 Model Analysis and Comparison

In this subsection, two extreme cases are considered to show the connections between our proposed heterogeneous graph information bottleneck (HGIB) and exiting methods.

Case 1: According to the decomposition in Eq. (6), the less the $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ in common, the more $I(\mathbf{v}^{(1)};\mathbf{h}^{(1)}|\mathbf{v}^{(2)})$ can be reduced by seeking the representation $\mathbf{h}^{(1)}$. Then, the more $I(\mathbf{v}^{(1)};\mathbf{h}^{(1)})$ can be used to discard specific information. Thus, the obtained node representation more robust. At the extreme, if $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ only share label information,

Datasets	Target classes	Node	Edges	Edge types	Features	Validation set	Training & Testing set	Meta-paths
ACM	3	8,994	25,922	4	1,902	300	2,725	PAP, PSP
DBLP	4	18,405	67,946	4	334	400	3,657	APA, APCPA
IMDB	3	12,772	37,288	4	1,256	300	2,639	MAM, MDM

Table 1: Statistics of the datasets

Datasets	Metrics	Train Ratio	Metapath2vec	GCN	GAT	DGI	HDGI	HAN	GTN	HGIB (random init)	HGIB (meta-path 1)	HGIB (meta-path 2)	HGIB
DBLP	Micro F1	20%	91.35	91.71	91.96	89.75	91.75	93.11	94.18	84.82	84.38	92.20	**95.09**
		40%	92.03	92.31	92.16	88.23	92.05	93.30	94.99	84.28	85.33	90.34	**96.13**
		60%	92.48	92.62	91.84	90.68	91.39	93.70	95.31	84.96	84.55	91.86	**96.20**
	Macro F1	20%	90.16	90.79	90.97	89.21	92.26	92.24	89.92	83.64	83.19	91.60	**94.76**
		40%	90.82	91.48	91.20	87.49	91.06	92.40	92.15	82.50	84.26	89.14	**94.89**
		60%	91.32	91.89	90.80	88.99	90.19	92.80	93.33	83.15	83.50	91.12	**94.52**
ACM	Micro F1	20%	65.00	86.77	86.01	91.04	92.27	89.22	92.68	84.95	88.94	69.45	**92.92**
		40%	69.75	87.64	86.79	92.23	91.45	89.64	**93.45**	84.34	88.13	69.39	92.80
		60%	71.29	88.12	87.40	92.36	91.31	89.33	91.09	85.32	88.35	68.35	**91.34**
	Macro F1	20%	65.09	86.81	86.23	91.04	92.32	89.40	90.44	84.93	88.96	67.70	**92.75**
		40%	69.93	87.68	87.04	92.23	90.01	89.79	89.53	84.42	88.16	67.56	**90.88**
		60%	71.47	88.10	87.56	92.36	90.83	89.51	89.42	85.36	88.37	65.10	**91.17**
IMDB	Micro F1	20%	45.65	49.78	55.28	57.28	58.93	55.73	60.92	52.32	58.57	57.48	**63.52**
		40%	48.24	51.71	55.91	58.16	59.94	57.97	60.96	52.96	58.21	57.41	**63.28**
		60%	49.09	52.29	56.44	58.95	**63.35**	58.32	61.39	52.46	58.90	56.43	61.57
	Macro F1	20%	41.16	45.73	49.44	56.90	**59.14**	50.00	55.17	23.27	46.19	42.80	58.23
		40%	44.22	48.01	50.64	57.23	58.09	52.71	58.51	23.08	39.58	36.94	**59.65**
		60%	45.11	49.15	51.90	56.35	**62.97**	54.24	61.10	23.12	46.77	33.32	60.22

Table 2: Comparison on node classification in terms of Micro F1 and Macro F1. Bold font indicates the best result. Note that both HAN and GTN are semi-supervised methods, while our proposed HGIB is an unsupervised one.

the node representation $\mathbf{h}^{(1)}$ is minimal for label $\mathbf{y}$, and self-supervised IB is equivalent to the supervised IB without accessing to the labels. Therefore, our proposed HGIB becomes the supervised heterogeneous graph neural network.

Case 2: In contrary, the more the $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ in common, the less $I(\mathbf{v^{(1)}}; \mathbf{h^{(1)}}|\mathbf{v^{(2)}})$ can be reduced by seeking the representation $\mathbf{h^{(1)}}$, then, the less $I(\mathbf{v^{(1)}}; \mathbf{h^{(1)}})$ can be reduced to discard specific information. Thus, the role of inducing multiple sub-graphs via different meta-paths is weakened. At the extreme, if $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ are identical, then the HGIB degenerates to the InfoMax principle [Hjelm *et al.*, 2019; Oord *et al.*, 2018] on homogeneous graph, i.e., maximizing $I(\mathbf{x}; \mathbf{h})$, such as in GMI [Peng *et al.*, 2020].

Information Theory based GNNs: Since the MINE [Belghazi *et al.*, 2018] and infoNCE [Oord *et al.*, 2018] were proposed, mutual information and contrastive learning are employed for self-supervised representation learning, such as infomax [Hjelm *et al.*, 2019]. Then, Deep Graph Informax (DGI) [Velickovic *et al.*, 2019] and InfoGraph [Sun *et al.*, 2020] are proposed for node and graph classification by applying MINE and infoNCE to graph data, respectively. HDGI [Park *et al.*, 2020] extends DGI to multiplex network by individually applying DGI to each graph and combining them via consistency regularization. Graph Information Bottleneck (GIB) [Wu *et al.*, 2020] is the first to incorporate IB [Tishby *et al.*, 2000] principle to GNNs. It main focus is the adversarial attack instead of performance improvement. Beside, GIB can't be directly applied to HIN to effectively exploit the rich information containing in multiple meta-paths.

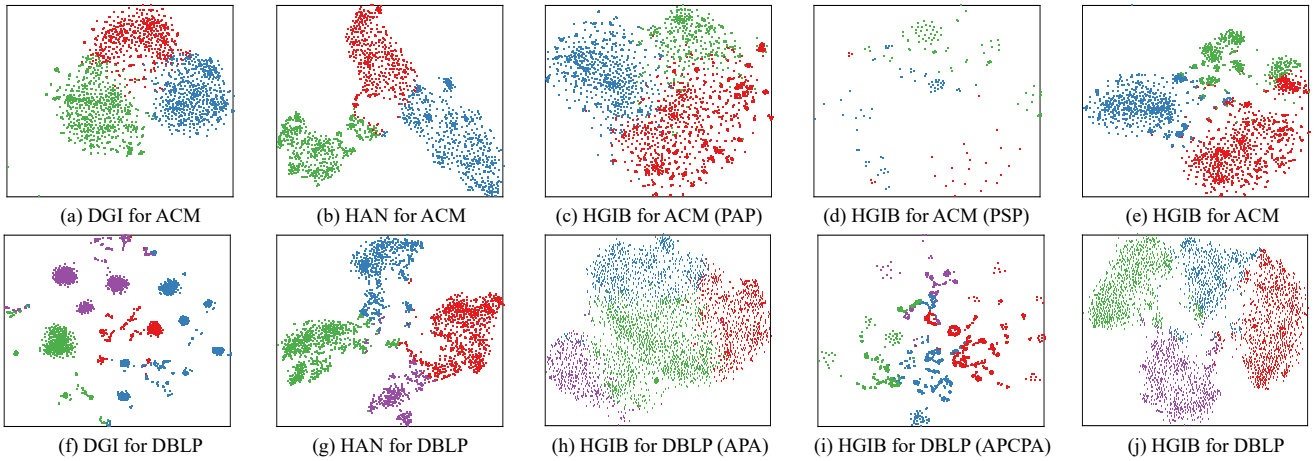
4 Experiments

In this part, we experimentally evaluate our proposed HGIB on node classification and clustering. HGIB firstly learn sets of node representations in unsupervised manner, then these representations are used for classification and clustering.

Datasets: Two citation network datasets ACM and DBLP, and a movie dataset IMDB, are employed for evaluation. The detailed descriptions of these heterogeneous graph data used in experiments are shown in Table 1.

Baselines: GHIB is compared with some state-of-the-art algorithms including traditional (heterogeneous) network embedding methods and graph neural networks, such as recent methods **DeepWalk** [Perozzi *et al.*, 2014] **metapath2vec** [Dong *et al.*, 2017], **GCN** [Kipf and Welling, 2017], **GAT** [Velickovic *et al.*, 2018], **HAN** [Wang *et al.*, 2019], **GTN** [Yun *et al.*, 2019], **DGI** [Velickovic *et al.*, 2019] and **HDGI** [Park *et al.*, 2020]. Note that, GCN, GAT, HAN and GTN are semi-supervised methods, while DGI, HDGI and our HGIB are unsupervised ones. Besides, the results obtained from HGIB with randomly initialization are provided to demonstrate the lifting power of information bottleneck.

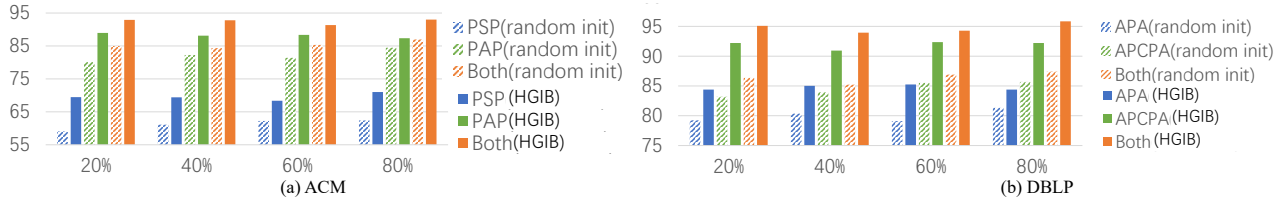
Implementation Details: The embedding dimension is set to 128 for all the above methods . Adam is employed as the optimizer. For IMDB and ACM, the learning rate is set to 10^{-3}, and the encoder consists of two layers GCN with the output dimensions of two GCN layers are 512 and 128 respectively. For DBLP, the learning rate is set as the same as above but the output dimensions of the GCN layer are 256 and 128 respectively. The early stopping with patience of 30



(a) DGI for ACM (b) HAN for ACM (c) HGIB for ACM (PAP) (d) HGIB for ACM (PSP) (e) HGIB for ACM

(f) DGI for DBLP (g) HAN for DBLP (h) HGIB for DBLP (APA) (i) HGIB for DBLP (APCPA) (j) HGIB for DBLP

Figure 2: The visualization of the embeddings obtained from DGI, HAN (semi-supervised) and our proposed HGIB on ACM and DBLP.



(a) ACM (b) DBLP

Figure 3: The performance gains of GNN trained with HGIB compared to the one with random initialization on ACM (a) and DBLP (b).

Methods	ACM		DBLP		IMDB	
Metrics	NMI	ARI	NMI	ARI	NMI	ARI
DeepWalk	41.61	35.10	76.53	81.35	1.45	2.15
metapath2	21.22	21.00	74.30	78.50	1.20	1.70
GCN	51.40	53.01	75.01	80.49	5.45	4.40
GAT	57.29	60.43	71.50	77.26	8.45	7.46
DGI	41.09	34.27	59.23	61.85	0.56	2.60
HDGI	54.35	49.48	60.76	62.27	1.87	3.70
HAN	61.56	64.39	79.12	84.76	10.87	10.01
GTN	55.98	49.62	65.86	63.49	4.32	1.90
HGIB	**70.55**	**71.47**	**88.46**	**87.29**	**12.09**	**10.80**

Table 3: Comparison on node clustering in ARI and NMI.

is utilized. The hyper-parameter γ is always fixed as 10^{-3}.

4.1 Classification and Clustering Results

For classification task, we randomly draw out 20%, 40% and 60% of nodes for training the linear regression classifier. This procedure is repeated for 20 times and report the averaged (Micro and Macro) F1-score. The results are shown in Table 2. It shows that HGIB achieves the best performance on almost all the datasets and various ratio of train set, especially for supervised methods such as GCN, GAT and HAN.

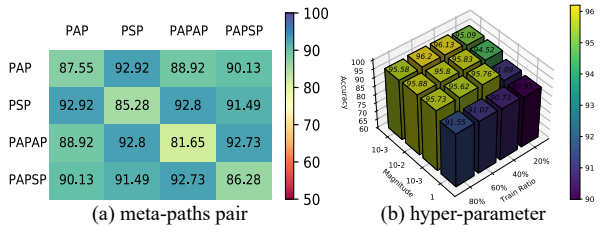
For clustering task, Kmeans is employed and the Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) are used to assess the quality of the clustering results. Similarly, the process is repeated for 20 times and the averaged results are reported in Table 3. HGIB significantly outperforms the baselines.

It can be observed that the performance improvements on clustering task are more remarkable compared to those on classification task. In fact, this phenomenon is not surprising due to the different focuses of clustering and classification tasks. Both of them are dependent on the comprehensiveness and importance of the representations. The classification task pays much attention on comprehensiveness, since the given labels play the role of feature selection, while the clustering pays much attention on the importance for the lack of supervision. Although HGIB enhances both comprehensiveness and importance of the representations, the importance improvement is more significant, because its intention is to discard the as much specific information as possible. Thus, its improvement on HGIB is more notable.

4.2 Case Study

Comparison with random parameters. To verify the improvements induced by the learned parameters via HGIB, the performances of learned parameters are compared with that of randomly initialized ones. The results are shown in Fig. 3. It shows that the performances are consistently and significantly improved, no matter the final representations are from one or two sub-graphs. This demonstrates the effectiveness of the proposed HGIB.

Impacts of pairs of meta-paths. The impacts of different pairs of meta-paths on performance are investigated. The results are shown in Fig. 4(a). It illustrates some interesting points. 1) The lowest performances are achieved when the two meta-paths are identical as shown on the diagonal. This



(a) meta-paths pair (b) hyper-parameter

Figure 4: Impacts of meta-paths pair (a) and hyper-parameter (b).

corresponds to the Case 2 in Sec 3.4. 2) If the two meta-paths are very different, the higher performance is achieved. For example, the best performance is achieved when author and subject appear in two meta-path, respectively, e.g. PAP vs. PSP. This corresponds to the Case 1 in Sec 3.4.

Hyper-parameter γ analysis. The hyper-parameter γ varies in 1, 10^{-1}, 10^{-2}, 10^{-3}. The impacts on performance on DBLP network with different presents of labels are shown in Fig. 4(b) and Tab. 4. Since the best performance is achieved when $\gamma = 10^{-3}$, thus this value is adopted as the default.

Datasets	Metrics	Train Ratio	1	10^{-1}	10^{-2}	10^{-3}
DBLP	Micro F1	20%	89.99	91.09	94.52	**95.09**
		40%	90.73	95.76	95.83	**96.13**
		60%	91.07	95.62	95.80	**96.20**
		80%	91.55	95.73	**95.88**	95.58
	Macro F1	20%	88.80	92.15	94.58	**94.76**
		40%	89.02	92.68	92.19	**94.89**
		60%	88.93	91.26	93.59	**94.52**
		80%	90.15	93.04	94.00	**94.06**
ACM	Micro F1	20%	86.37	90.07	**93.38**	92.92
		40%	87.81	91.76	91.75	**92.80**
		60%	87.93	89.81	90.05	**91.34**
		80%	89.26	90.63	**93.79**	93.45
	Macro F1	20%	82.55	89.97	92.36	**92.75**
		40%	84.60	90.39	**91.46**	90.88
		60%	85.53	90.04	90.95	**91.17**
		80%	87.48	91.00	90.47	**92.62**
IMDB	Micro F1	20%	54.18	60.78	62.95	**63.52**
		40%	56.98	61.10	60.41	**63.28**
		60%	56.84	60.30	61.18	**61.57**
		80%	53.29	60.05	**60.17**	54.44
	Macro F1	20%	49.77	56.02	56.10	**58.23**
		40%	50.09	58.62	59.44	**59.65**
		60%	48.77	53.64	59.86	**60.22**
		80%	44.93	58.25	57.46	**60.20**

Table 4: Impact of Hyper-parameter γ

5 Conclusions

This paper replaces the complementary hypothesis of the multiple homogeneous attributed networks induced by different meta-paths, which has been Heterogeneous Graph Neural Networks, with an alternative consensus one. The consensus assumption is implemented as Heterogeneous Graph Information Bottleneck (HGIB) by extending information bottleneck to unsupervised manner via self-supervised strategy. The proposed HGIB can be regard as the generalizations of both the semi-supervised heterogeneous GNNs and the infomax on homogeneous graph. Extensive experiments on real datasets demonstrate the correctness and the effectiveness of the consensus hypothesis by verifying the significant performance improvement of the unsupervised HGIB compared to the most semi-supervised state-of-the-art methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61972442, Grant U1936208 and Grant 61802282, in part by the Key Research and Development Project of Hebei Province of China under Grant 20350802D and 20310802D; in part by the Natural Science Foundation of Hebei Province of China under Grant F2020202040, in part by the Hebei Province Innovation Capacity Enhancement Project under Grant 199676146H, in part by the Natural Science Foundation of Tianjin of China under Grant 20JCYBJC00650, and in part by the Key Program of the Chinese Academy of Sciences under Grant QYZDB-SSW-JSC003.

A Proof of Eq. (10)

$$
\begin{aligned}
& I_{\Theta^{(1)}}(\mathbf{v^{(1)}};\mathbf{h^{(1)}}|\mathbf{v^{(2)}}) \\
= \;& \mathbb{E}_{v_1,v_2}\mathbb{E}_h\left[\log\frac{p_{\Theta^{(1)}}(h^{(1)}=h|v^{(1)}=v_1)}{p_{\Theta^{(1)}}(h^{(1)}=h|v^{(1)}=v_2)}\right] \\
= \;& \mathbb{E}_{v_1,v_2}\mathbb{E}_h\left[\log\frac{p_{\Theta^{(1)}}(h^{(1)}=h|v^{(1)}=v_1)}{p_{\Theta^{(2)}}(h^{(2)}=h|v^{(2)}=v_2)}\right] \\
& +\mathbb{E}_{v_1,v_2}\mathbb{E}_h\left[\log\frac{p_{\Theta^{(2)}}(h^{(2)}=h|v^{(2)}=v_2)}{p_{\Theta^{(1)}}(h^{(1)}=h|v^{(1)}=v_2)}\right] \\
= \;& D_{KL}(p_{\Theta^{(1)}}(h^{(1)}|v^{(1)})||p_{\Theta^{(2)}}(h^{(2)}|v^{(2)})) \\
& -D_{KL}(p_{\Theta^{(1)}}(h^{(2)}|v^{(1)})||p_{\Theta^{(2)}}(h^{(2)}|v^{(2)})) \\
\leq \;& D_{KL}(p_{\Theta^{(1)}}(\mathbf{h^{(1)}}|\mathbf{v^{(1)}})||p_{\Theta^{(2)}}(\mathbf{h^{(2)}}|\mathbf{v^{(2)}})), \qquad (16)
\end{aligned}
$$

B Proof of Eq. (12)

$$
\begin{aligned}
& I_{\Theta^{(1)}}(\mathbf{v^{(2)}};\mathbf{h^{(1)}}) \\
= \;& I_{\Theta^{(1)}\Theta^{(2)}}(\mathbf{h^{(1)}};\mathbf{h^{(2)}v^{(2)}}) - I_{\Theta^{(1)}\Theta^{(2)}}(\mathbf{h^{(1)}};\mathbf{h^{(2)}}|\mathbf{v^{(2)}}) \\
= \;& I_{\Theta^{(1)}\Theta^{(2)}}(\mathbf{h^{(1)}};\mathbf{h^{(2)}v^{(2)}}) \\
= \;& I_{\Theta^{(1)}\Theta^{(2)}}(\mathbf{h^{(1)}};\mathbf{h^{(2)}}) + I_{\Theta^{(1)}\Theta^{(2)}}(\mathbf{h^{(1)}};\mathbf{v^{(2)}}|\mathbf{h^{(2)}}) \\
\geq \;& I_{\Theta^{(1)}\Theta^{(2)}}(\mathbf{h^{(1)}};\mathbf{h^{(2)}}), \qquad (17)
\end{aligned}
$$

References

[Achille and Soatto, 2018] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *JMLR*, 19:50:1–50:34, 2018.

[Alemi *et al.*, 2017] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017.

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio,

R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *ICML*, pages 530–539. PMLR, 2018.

[Bruna *et al.*, 2014] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. In *ICLR*, 2014.

[Dong *et al.*, 2017] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *SIGKDD*, pages 135–144. ACM, 2017.

[Federici *et al.*, 2020] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2020.

[Fu *et al.*, 2020] Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. MAGNN: metapath aggregated graph neural network for heterogeneous graph embedding. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW*, pages 2331–2341. ACM / IW3C2, 2020.

[Hjelm *et al.*, 2019] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

[Hu *et al.*, 2019] Linmei Hu, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. Heterogeneous graph attention networks for semi-supervised short text classification. In *EMNLP-IJCNLP*, pages 4820–4829, 2019.

[Hu *et al.*, 2020] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW*, pages 2704–2710. ACM / IW3C2, 2020.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[Li *et al.*, 2018] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI*, pages 3538–3545, 2018.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Park *et al.*, 2020] Chanyoung Park, Donghyun Kim, Jiawei Han, and Hwanjo Yu. Unsupervised attributed multiplex network embedding. In *AAAI*, pages 5371–5378, 2020.

[Peng *et al.*, 2020] Zhen Peng, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang. Graph representation learning via graphical mutual information maximization. In *WWW*, pages 259–270, 2020.

[Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In *SIGKDD*, pages 701–710. ACM, 2014.

[Shi *et al.*, 2017] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and Philip S. Yu. A survey of heterogeneous information network analysis. *TKDE*, 29(1):17–37, 2017.

[Shi *et al.*, 2019] Chuan Shi, Binbin Hu, Wayne Xin Zhao, and Philip S. Yu. Heterogeneous information network embedding for recommendation. *TKDE*, 31(2):357–370, 2019.

[Sun and Han, 2012] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: a structural analysis approach. *SIGKDD Explor.*, 14(2):20–28, 2012.

[Sun *et al.*, 2020] Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*, 2020.

[Tishby *et al.*, 2000] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[Velickovic *et al.*, 2018] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[Velickovic *et al.*, 2019] Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *ICLR*, 2019.

[Wang *et al.*, 2019] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. Heterogeneous graph attention network. In *WWW*, pages 2022–2032. ACM, 2019.

[Wang *et al.*, 2020] Xiao Wang, Deyu Bo, Chuan Shi, Shaohua Fan, Yanfang Ye, and Philip S Yu. A survey on heterogeneous graph embedding: Methods, techniques, applications and sources. *arXiv preprint arXiv:2011.14867*, 2020.

[Wu *et al.*, 2019] Felix Wu, Amauri H. Souza Jr., Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *ICML*, pages 6861–6871, 2019.

[Wu *et al.*, 2020] Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. Graph information bottleneck. In *NeurIPS*, 2020.

[Wu *et al.*, 2021] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *TNNLS*, 32(1):4–24, 2021.

[Yun *et al.*, 2019] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. Graph transformer networks. In *NeurIPS*, pages 11960–11970, 2019.