GRAPH FINE-GRAINED CONTRASTIVE REPRESENTATION LEARNING

Hui Tang, Xun Liang(✉), Yuhui Guo , Xiangping Zheng,Bo Wu

School of Information, Renmin University of China

ABSTRACT

Existing graph contrastive methods have benefited from ingenious data augmantations and mutual information estimation operations that are carefully designated to augment graph views and maximize the agreement between representations produced at the aftermost layer of two view networks. However, the design of graph CL schemes is coarse-grained and difficult to capture the universal and intrinsic properties across intermediate layers. To address this problem, we propose a novel fine-grained graph contrastive learning model (FGCL), which decomposes graph CL into global-to-local levels and disentangles the two graph views into hierarchical graphs by pooling operation to capture both global and local dependencies across views and across layers. To prevent layers mismatch and automatically assign proper hierarchical representations of the augmented graph (Key view) for each pooling layer of the original graph ($Query$ view), we propose a sematic-aware layer allocation strategy to integrate positive guidance from diverse representations rather than a fixed layer manually. Experimental results demonstrate the advantages of our model on graph classification task. This suggests that the proposed fine-grained graph CL presents great potential for graph representation learning.

***Index Terms*—** Contrastive learning, Graph representation, Graph pooling, Graph augmentation, Sematic allocation

1. INTRODUCTION

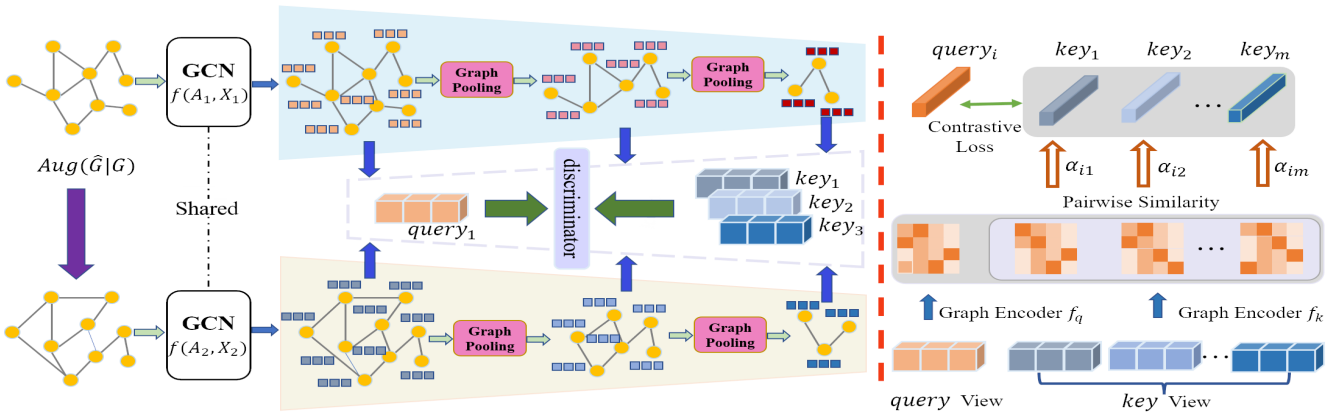
Graph Neural Networks (GNNs) has emerged as a powerful tool for analyzing graph related tasks, such as node classification [1], graph classification [2] and link prediction [3]. However, existing GNN models are mostly trained under supervision and require abundant labeled nodes. Contrastive learning (CL) as an important renaissance member of self-supervised learning (SSL), reduces the dependency on excessive annotated labels and achieves great success in many fields. These CL methods leverage the classical Information Maximization principle and seek to maximize the Mutual Information (MI) by contrasting positive and negative pairs.

This work was supported by the National Natural Science Foundation of China (62072463, 71531012), and the National Social Science Foundation of China (18ZDA309). Xun Liang is the corresponding author of this paper.

For graph CL, data augmentation and mutual information estimation operations, proved to be critical components. Graph data augmentations are broadlly classified into four types: Node dropping [4], Edge perturbation [5], Attribute masking [6] and Subgraph sampling [7], each of which imposes certain prior over graph data and parameterized for the extent and pattern. MI estimation mainly quantifies the correlation between the latent representations of two graph views. The representations include node representations and corresponding high-level summaries of graphs and so the MI estimation objects can be divided into node-level vs node-level, node-level vs graph-level and graph-level vs graph-level.

The node-level focuses on local representation patterns. GCA [8] introduced a adaptive data augmentation to contrast at the node level. GMI [9] extents the idea of GCA to a form of weighted sum through cross-layer node contrasting. The graph-level plays an important role to preserve the global dependencies. GraphCL [10] applies a series of graph augmentations and then learns to predict whether two graphs originate from the same graph or not. GCC [11] first samples multiple subgraphs to capture the universal graph topological properties across multiple graphs. Different from discriminating only at node-level or graph-level, maximizing MI between global-local representations can preserve dependencies. DGI [12] followed on the idea of DIM [13] and proposed to contrast the high-level patch representations. MVGRL [14] maximized the MI between the cross-view representations of nodes and large-scale graphs. Although these graph CL approaches leveraged different MI estimation objects, the design of graph CL schemes is coarse-grained and ignore a wealth of information contained in the intermediate layers.

In this work, we propose a fine-grained graph contrastive learning model, which decomposes graph CL into global-to-local levels to capture both global and local dependencies instead of contrasting between input graph and the representations that generated at the aftermost layer of network or through stacking multiple layers without selection. The proposed FGCL applies the idea of dictionary query to match intermediate representations between the original graph ($query$ view) and the augmented graph (key view) to compute local mutual information across views and across layers. We choose edge diffusion based on structure enhancement [14] to transform a $query$ view into a correlated key view. We follow the graph pooling method based on self-attention [15] to



(a) Fine-grained contrast cross views and layers (b) Semantic-aware layer allocation

Fig. 1. An architecture of the proposed fine-grained graph contrastive learning (FGCL) model.

consider both node features and graph topology. To prevent layers mismatch and automatically assign proper hierarchical representations of the key view for each pooling layer of the query view, we propose a semantic-aware layer allocation strategy to integrate positive guidance from diverse representations rather than a fixed layer manually.

2. FINE-GRAINED CONTRATIVE LEARNING METHOD ON GRAPHS

2.1. Semantic-aware Layers Allocation

Our approach applies the idea of dictionary query to match intermediate representations between the original graph ($Query$ view) and the augmented graph (Key view) to compute local mutual information across views and across layers. Firstly, we employ the data augmentation based on edge-diffusion to produce a Key graph view by removing some unreliable connections in the complex input graph. Secondly, due to current GCNs are inherently flat and do not learn hierarchical representations of graphs, which have limits to capture the structural patterns behind these graphs, we propose to use SAGPool as our pooling function to obtain effective represention hierarchically. Then the sets of intermidiare representations generated by pooling layers for $Query$ and Key views are:

$$(F_1, F_2, ..., F_Q) = P_1(G), F_q \in R^{N \times f^q} \quad (1)$$

$$(\hat{F_1}, \hat{F_2}, ..., \hat{F_K}) = P_2(\hat{G}), \hat{F_k} \in R^{N \times \hat{f^k}} \quad (2)$$

P_1 and P_2 is SAGPool function with Q and K pool layers. F_q and $\hat{F_k}$ are graph-level representations, which can be summaried from node-level representations.

From a dictionary look-up perspective, finding a optimal layer association set cross views and cross layers is crucial for implementing our fine-grained graph CL. The layer association sets $\mathcal{C}$ of existing methods are generated by random selection or one-to-one match, which is simple and lead to loss useful information due to semantic level of those intermediates vary among two views.

Rather than performing MI estimation based on fixed associations, we propose a semantic-aware layer allocation to capture global and local dependencies. Each F_q in $Query$ view is automatically associated with those semantic-related target $\hat{F_k}$ in Key view by attention allocation, as illustrated in Figure 1. The learned association set $\mathcal{C}$ is denoted as

$$\mathcal{C} = \{(F_q, \hat{F_k}), \forall q \in [1, ..., Q], \forall k \in [1, ..., K]\} \quad (3)$$

Due to inconsistent dimensions across layers, we need to project the representations of each $\hat{F}$ into the $query$ view to align with the spatial dimension of each query layer. The projection operation can be representated as

$$\theta(F, \hat{F}) = Proj(\hat{F} \in R^{N \times K \times \hat{f}} \mapsto R^{N \times K \times f}) \quad (4)$$

Each function $Proj(.)$ includes two MLP encoders that encode the query instance F_q and each key instance $\hat{F_k}$ to the same dimensional representations. After that, to measure the inherent semantic association of intermediate layers by similarity matrices. Firstly, we use $L2$-normalization to normalize the projection matrix or two views as Z_F and $Z_{\hat{F}}$. Then, the two normalized matrix can be used to capture the similarity of N graphs as S_F and $S_{\hat{F}}$ as follows:

$$S_F = Z_F \times Z_F^T \qquad S_{\hat{F}} = Z_{\hat{F}} \times Z_{\hat{F}}^T \quad (5)$$

where $S_F \in R^{N \times Q}$ and $S_{\hat{F}} \in R^{N \times K}$ matrices. Q and K are the number of pooling layers in the two views.

Based on the self-attention framework, we separately encode the pairwise similarity matrices of each query layer and key layer into two subspaces by a shared MLP layer to alleviate the effect of noise and sparseness. Here, we focus on a graph $\mathcal{G}_i$, where its similarity vetors in S_F and $S_{\hat{F}}$ are

$S_F[i] \in R^{1\times Q}$, $S_{\hat{F}}[i] \in R^{1\times K}$ respectively. The embeddings generated by MLP are

$$Q_F[i] = MLP(S_F[i]) \qquad Q_{\hat{F}}[i] = MLP(S_{\hat{F}}[i]) \tag{6}$$

The parameters of $MLP(.)$ learned during training to generate query and key vectors and shared by all instances. Then, the attention values $\alpha^i_{(F_q,\hat{F}_k)}$ is calculated as follows

$$\alpha^i_{(F_q,\hat{F}_k)} = \frac{Q_{F_q}[i] \times Q^T_{\hat{F}_k}[i]}{\sum_{k=1}^{K} Q_{F_q}[i] \times Q^T_{\hat{F}_k}[i]} \tag{7}$$

with the corresponding weight satisfies $\sum_{k=1}^{K} \alpha^i_{(F_q,\hat{F}_k)} = 1, \forall q \in [1, ..., Q]$. The weight $\alpha^i_{(F_q,\hat{F}_k)} \in R^{Q\times K}$ represents the extent to which the $query$ layer q is attended in deriving the semantic-aware guidance for the key layer k. Attention-based allocation provides a possible way to suppress negative effects caused by layer mismatch and integrate positive guidance from multiple target layers.

2.2. Mutual Information Maximization

Our proposed FGCL model can be regarded as fine-grained contrastive learning framework, which decomposes graph CL into global-to-local levels to caputer global and universal dependencies cross views and cross layers. For any graph $\mathcal{G}_i$, its q-th layer represention generated in the $Query$ view, F_q, is treated as the anchor, the represention of it generated in the $kKey$ view, $\hat{F}_k$ forms the positive sample, and the other representations of $\mathcal{G}_{j(\neq i)}$ in the two views are naturally regarded as negative samples.

In our fine-grained graph CL setting, we define the pairwise objective for each sampled pair $(F_q, \hat{F}_k)$. The distribution of F_q and $\hat{F}_k$ are $P_{q^*} = P(q^* = f_q(F_q))$ and $P_{k^*} = P(k^* = f_k(\hat{F}_k))$. The mutual information between $query$ and key views is computed as the KL-divergence between the joint distribution P_{q^*,k^*} of P_{q^*} and P_{k^*}, and the product of marginal distributions $P_{q^*} \otimes P_{k^*}$:

$$\begin{aligned} MI(q,k) &= D_{KL}(P_{q^*,k^*} || P_{q^*} \otimes P_{k^*}) \\ &\geq \sup_{T\in\mathcal{T}} \{E_{P_{(q^*,k^*)}} \log \sigma([T(F_q, \hat{F}_k)]) \\ &\quad - E_{(P_{q^*},P_{k^*})} \log\left(1 - \sigma([T(F_q, \hat{F}_k)])\right)\} \end{aligned} \tag{8}$$

where $MI(q,k)$ is the mutual information between output of q-th layer and k-layer in the $Query$ and Key views respectively, and its lower bound is learned via contrastive learning. σ is the sigmoid function. $T \in \mathcal{T}$ is an arbitrary function that maps representations a pair of $(F_q, \hat{F}_k)$ to a real value, reflecting their dependencies. We set T is two layers MLP (Multi-Layer Perceptron) with 128, 64 neurons respectivly. Of course, more complex functions also can be considerd if the accuracy and complexity of the model is acceptable.

After dimensional projections and semantic-aware layers allocation of each pooling layer, we maximize the mutual information of intermidiate representations generated by hierarchical pooling operation through cross views and cross layers. For a mini-batch graphs with size N, the $query$ view produces Q represention and the key view produces K represention. The mutual information loss between their representations of the two view is

$$L_{MI} = \sum_{q=1}^{Q}\sum_{k=1}^{K}\sum_{i=1}^{N} \alpha^i_{(F_q,\hat{F}_k)} MI^i(q,k) \tag{9}$$

3. EXPERIMENT AND ANALYSIS

In this section, we conduct experiments to assess and rationalize our model through answering the following questions.

- **RQ1**: Does our proposed FGCL model outperform existing baseline methods on graph classification?
- **RQ2**: Does the proposed sematic-aware layer allocation scheme benefit the learning of FGCL?
- **RQ3**: Is the proposed model sensitive to hyperparameters of layer number in the $Query - Key$ views?

We use eight datasets from TUDataset to evaluate our proposed FGCL model, including three social network, three bioinformatics and two molecules graph datasets. Table 1 summarizes the statistics of these datasets. We compare FGCL with several recent developed graph classification models. Based on the MI estimation objects, we divide these methods into three types: node vs node (GMI [9], GCA [8] and GIC [16]), node vs graph (DGI [12], MVGRL [14], HDMI [17] and DMGI [18]) and graph vs graph (GraphCL [10], GCC [11], InfoGraph [19] and SUGR [20]).

3.1. Comparison with State-of-the-Art (RQ1)

Table 1 summarizes the graph classification accuracy (%) of FGCL and eleven different baselines. In general, it can be seen from the table that our proposed model shows strong performance on all datasets. Specifically, FGCL outperforms state-of-the-art models by up to 1.2%, 3.08%, 3.35%, 2.3%, 2.47%, 1.94% and 2.37% on COLLAB, IMDB-MULTI, DD, ENZYMES, PROTEINS, NCI1 and Mutagenicity respectively. Moreover, we make other observations as follows. FGCL achieves state-of-the-art performance on all datasets except the DBLP dataset, where the accuracy of FGCL is only lower than that of 0.96% the best baseline. The reason may be that the number of edges in the dataset is relatively sparse, which the graph structure is relatively clean, and there is relatively little noise injected. In summary, the superior performance of FGCL compared to existing state-of-the-art methods that fine-grained graph CL performs well to the graph representation learning.

Table 1. Summary of datasets statistics and Graph classification accuracies (%) of different methods.

Dataset		Social Networks			Bioinformatics			Small Molecules	
Methods		COLLAB	IMDB-MULTI	DBLP	D&D	ENZYMES	PROTEINS	NCI1	Mutagenicity
# Graphs		5000	1500	19456	1178	600	1113	4110	4337
# Classes		3	3	2	2	6	2	2	2
Avg # Nodes		74.49	13.00	10.48	284.32	32.63	39.06	29.87	30.32
Avg # Edges		2457.78	65.94	19.65	715.66	62.14	72.82	32.30	30.77
Node vs Node	GMI	76.25±1.58	50.09±1.21	79.27±1.26	76.59±0.69	56.00±2.76	70.24±2.47	73.54±0.77	76.62±1.07
	GCA	74.74±1.97	51.43±0.83	81.70±2.27	79.20±0.74	49.83±2.57	69.58±1.96	74.63±0.96	77.96±0.85
	GIC	75.39±1.26	49.38±1.07	**83.63±1.93**	82.69±0.70	54.60±2.35	71.33±2.53	74.59±0.85	79.04±0.93
Node vs Graph	DGI	73.90±0.99	51.77±0.96	78.43±1.85	75.05±0.63	48.91±2.41	71.44±2.01	76.87±0.93	78.43±0.95
	MVGRL	76.85±1.46	52.18±0.78	77.65±2.36	81.94±0.71	50.58±2.47	69.89±2.37	75.17±0.70	78.94±0.86
	HDMI	74.08±1.38	50.46±0.89	80.90±2.50	82.43±0.65	55.09±2.93	68.42±1.88	74.67±0.82	79.20±1.56
	DMGI	75.34±1.55	49.89±0.83	81.48±2.22	81.67±0.62	54.19±1.84	72.09±2.45	76.43±0.69	77.40±0.84
Graph vs Graph	GraphCL	77.56±1.49	50.27±0.67	77.06±1.97	81.23±0.56	50.35±2.16	70.70±1.68	75.41±0.95	76.88±1.43
	GCC	73.34±1.63	48.76±0.74	81.94±1.56	80.16±0.74	51.03±2.43	71.33±2.18	74.36±1.00	77.91±1.25
	InfoGraph	75.49±1.80	52.91±0.99	80.35±2.03	82.90±0.63	53.41±2.34	69.46±1.73	75.87±0.94	77.35±0.87
	SUGR	76.20±0.80	50.14±0.83	82.59±2.56	80.44±0.80	52.88±2.29	71.36±1.84	77.52±0.78	78.99±1.00
Our model		**78.76±1.78**	**54.85±0.75**	82.67±2.35	**85.29±0.76**	**57.39±1.67**	**73.91±1.88**	**79.46±0.91**	**81.57±1.01**

3.2. Semantic-aware Layer Allocation Analysis (RQ2)

The section analysis the impact of layer allocation strategy. There are two simple allocation strategies, random selection and one-to-one match. FGCL$_{random}$ denotes the model with a uniform $Bernoulli$ function for random selection. The variant FGCL$_{oto}$ is defined as a fixed match mode through layer by layer sequentially. The results are presented in Table 2, where we can see that FGCL equiped with the semantic-aware layer allocation scheme improves model performance consistently on all datasets. Specially, on the COLLAB dataset, our proposed FGCL gains 5.46% absolute improvement compared to the simple match methods of random selection and one-to-one, which reveals that the semantic-aware layers allocation strategy based on the attention mechanism can prevent the negative effects of exsting mismatched layers and integrate positive guidance from multiple target layers to capture both global and local dependencies.

Table 2. Results of FGCL and its two variants.

Architecture	COLLAB	D&D	NCI1
FGCL$_{random}$	74.75±1.83	82.46±0.69	78.16±0.99
FGCL$_{oto}$	73.30±2.06	81.23±0.79	76.09±1.32
FGCL	**78.76±1.78**	**85.29±0.76**	**79.46±0.91**

3.3. Sensitivity Analysis (RQ3)

We study two important hyperparameters, the number of layers Q and K in the $Query$ and Key views. For sake of visualization brevity, we vary these hyperparameters from 1 to 6. The results on the COLLAB dataset are shown in Figure 2. It can be observed that the classification accuray is relatively stable when the parameters are not too large. However, when the layer number exceeds 4, the model performance will be heavily undermined. For example when Q=6, no matter how the K value changes, the accuacy is very poor. We conjecture that the representations will converge to a stationary point and lead to vanishing gradients with layer depth increasing. Another interesting finding is that the model performance becomes better as the increase of K value within a reasonable range, which indicates that more valuable information from the augmented view pooled by multiple layers is participated in the fine-grained comparison process.

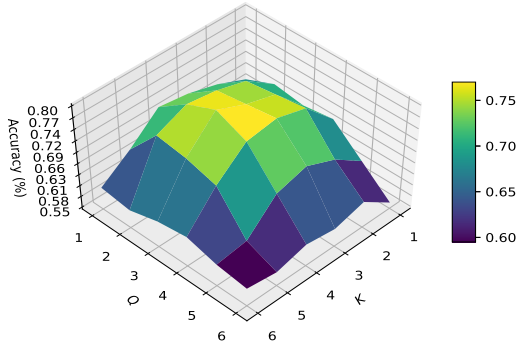


Fig. 2. The hyperparameters analysis of Q and K.

4. CONCLUSION

In this work, we present FGCL model, which was a novel fine-grained graph contrastive learning framework to disentangle the two graph views into hierarchical representations and decompose contrastive learning into global-to-local levels across views and across layers. The sematic-aware layer allocation strategy was proposed to integrate positive guidance from diverse representations rather than a fixed layer manually to capture the global and local dependencies well.

5. REFERENCES

[1] Kangjie Li, Yixiong Feng, Yicong Gao, and Jian Qiu, "Hierarchical graph attention networks for semi-supervised node classification," *Applied Intelligence*, vol. 50, no. 10, pp. 3441–3451, 2020.

[2] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen, "An end-to-end deep learning architecture for graph classification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[3] Lei Cai and Shuiwang Ji, "A multi-scale approach for graph link prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 3308–3315.

[4] Wenzheng Feng, Jie Zhang, Yuxiao Dong, Yu Han, Huanbo Luan, Qian Xu, Qiang Yang, Evgeny Kharlamov, and Jie Tang, "Graph random neural network for semi-supervised learning on graphs," *arXiv preprint arXiv:2005.11079*, 2020.

[5] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang, "Dropedge: Towards deep graph convolutional networks on node classification," *arXiv preprint arXiv:1907.10903*, 2019.

[6] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec, "Strategies for pre-training graph neural networks," *arXiv preprint arXiv:1905.12265*, 2019.

[7] Yizhu Jiao, Yun Xiong, Jiawei Zhang, Yao Zhang, Tianqi Zhang, and Yangyong Zhu, "Sub-graph contrast for scalable self-supervised graph representation learning," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 222–231.

[8] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang, "Graph contrastive learning with adaptive augmentation," in *Proceedings of the Web Conference 2021*, 2021, pp. 2069–2080.

[9] Zhen Peng, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang, "Graph representation learning via graphical mutual information maximization," in *Proceedings of The Web Conference 2020*, 2020, pp. 259–270.

[10] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen, "Graph contrastive learning with augmentations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5812–5823, 2020.

[11] Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang, "Gcc: Graph contrastive coding for graph neural network pre-training," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1150–1160.

[12] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm, "Deep graph infomax," *arXiv preprint arXiv:1809.10341*, 2018.

[13] R Devon Hjelm, Alex Fedorov, and Samuel Lavoie-Marchildon, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

[14] Kaveh Hassani and Amir Hosein Khasahmadi, "Contrastive multi-view representation learning on graphs," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4116–4126.

[15] Junhyun Lee, Inyeop Lee, and Jaewoo Kang, "Self-attention graph pooling," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3734–3743.

[16] Costas Mavromatis and George Karypis, "Graph infoclust: Leveraging cluster-level node information for unsupervised graph representation learning," *arXiv preprint arXiv:2009.06946*, 2020.

[17] Baoyu Jing, Chanyoung Park, and Hanghang Tong, "Hdmi: High-order deep multiplex infomax," in *Proceedings of the Web Conference 2021*, 2021, pp. 2414–2424.

[18] Chanyoung Park, Donghyun Kim, Jiawei Han, and Hwanjo Yu, "Unsupervised attributed multiplex network embedding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 5371–5378.

[19] Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang, "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization," *arXiv preprint arXiv:1908.01000*, 2019.

[20] Qingyun Sun, Jianxin Li, Hao Peng, and Jia Wu, "Sugar: Subgraph neural network with reinforcement pooling and self-supervised mutual information mechanism," in *Proceedings of the Web Conference 2021*, 2021, pp. 2081–2091.