Disentangled Graph Self-supervised Learning for Out-of-Distribution Generalization

Haoyang Li [1] **Xin Wang** [1] **Zeyang Zhang** [1] **Haibo Chen** [1] **Ziwei Zhang** [1] **Wenwu Zhu** [1]

Abstract

Graph out-of-distribution (OOD) generalization, aiming to generalize graph neural networks (GNNs) under distribution shifts between training and testing environments, has attracted ever-increasing attention recently. However, existing literature heavily relies on sufficient task-dependent graph labels, which are often scarce or even unavailable, limiting their applications in real-world scenarios. In this paper, we study the self-supervised graph OOD generalization problem, *i.e.*, learning GNNs capable of achieving relatively stable performances under distribution shifts without graph labels. However, the problem remains largely unexplored, with the critical challenge that the invariant and variant information are highly entangled in graphs. To solve this problem, we propose an OOD generalized disentangled graph contrastive learning model (**OOD-GCL**), which is capable of learning disentangled graph-level representations with self-supervision that can handle distribution shifts between training and testing graph data. Specifically, we first introduce a disentangled graph encoder to map each input graph into the factorized graph representation. Then we propose a tailored disentangled invariant self-supervised learning module to maximize predictive ability of the representations and make sure the representations other than from one specific channel are invariant to the environments partitioned by this latent factor for excluding the information corresponding to this latent factor for disentanglement. Finally, the disentangled graph representations are fed into a linear predictor and finetuned for the downstream tasks. We provide comprehensive theoretical analyses to show that our model can learn disentangled graph representations and achieve OOD generalization. Extensive experiments on real-world datasets demonstrate the superiority of our model against state-of-the-art baselines under distribution shifts for graph classification tasks.

1. Introduction

Graph structured data is ubiquitous in the real world, such as social networks (Qiu et al., 2018), traffic networks (Yu et al., 2017), financial networks (Yang et al., 2021), chemical molecules (Hu et al., 2020), etc. In the last decade, graph neural networks (GNNs) (Kipf & Welling, 2017; Veličković et al., 2018; Xu et al., 2019) have been a central topic in graph machine learning and made great progress in both academia and industry. Most of the existing literature is built upon the in-distribution (I.D.) hypothesis that assumes the testing and training graph data are from the identical distribution. However, in the real world, distribution shifts between testing and training graphs widely and inevitably exist, where the performance of existing GNNs drops significantly due to lacking out-of-distribution (OOD) generalization capacity (Li et al., 2022b).

Despite the notable success of graph OOD generalization methods, the existing literature relies on sufficient task-dependent annotated labels to learn OOD generalized graph representations, which could be extremely scarce, or even unavailable in practice. For example, in the context of drug discovery, it necessitates high costs and substantial human labor in clinical tests to obtain labeled data for training OOD generalized graph models in a supervised manner, if we want to predict the invariant properties of molecules under distribution shifts (Paul et al., 2021). Since the existing approaches heavily rely on supervised labels, they will fail to learn invariant representations and model truly predictive relations with labels under distribution shifts for OOD generalization, when graph labels are scarce or not available, leading to severe performance degeneration.

In this paper, we study the self-supervised graph OOD generalization problem, i.e., learning GNNs capable of achieving relatively stable performances under distribution shifts

[1] Department of Computer Science and Technology, BNRIST, Tsinghua University, Beijing, China. Correspondence to: Xin Wang <xin_wang@tsinghua.edu.cn>, Wenwu Zhu <wwzhu@tsinghua.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

without graph labels. The key goal is to pretrain the graph encoder with much more available unlabeled graph data in a self-supervised manner to produce disentangled graph representations that can achieve promising OOD generalization when evaluated on testing graphs under distribution shifts. However, self-supervised graph OOD generalization is non-trivial due to the following challenges. First, it is challenging to get rid of spurious correlations between invariant and variant information in graph representations. The existing graph self-supervised methods generally adopt a holistic view meaning that the learned representations describe graphs as a perceptual whole (Li et al., 2021a), so that the invariant and variant information are highly entangled in the graph representations. Second, the existing methods tend to conduct pretraining tasks with spurious variant information (Li et al., 2022e), which fail to capture disentangled factors that reflect the semantic information which can be the labels in downstream tasks, leading to poor generalization performances.

To tackle these challenges, we propose an OOD generalized disentangled graph contrastive model (**OOD-GCL**), which is capable of learning disentangled graph-level representations with self-supervision that can handle distribution shifts between training and testing graph data. Specifically, we first introduce a disentangled graph encoder to map each input graph into the corresponding factorized graph representation. Each channel is specifically designed to capture features from a specific disentangled latent factor, thereby accurately representing the information of that disentangled latent factor. Then we propose a tailored disentangled invariant self-supervised learning module. It can achieve sufficient predictive ability of the representations and make sure the representations other than from one specific channel are invariant to the environments partitioned by this latent factor for excluding the information corresponding to this latent factor. Thus the representations are encouraged to be disentangled so as to best characterize the aspect pertinent to a latent factor of the graph and achieve OOD generalized predictions when finetuning in downstream tasks. We further provide comprehensive theoretical analyses to show that our proposed model can learn disentangled graph representations and achieve OOD generalization with a strong guarantee. Extensive empirical evaluations are performed on several well-established graph benchmarks. The results demonstrate that the representations generated by our model lead to significant enhancements in generalization performance for downstream graph classification tasks under distribution shifts, outperforming the state-of-the-art baselines.

The contributions of this paper are summarized as follows.

- We focus on a novel self-supervised graph out-of-distribution (OOD) generalization problem and propose a tailored model to learn disentangled graph-level representations with self-supervision that can handle distribution shifts between training and testing graph data. To the best of our knowledge, we are the first to study self-supervised graph OOD generalization with theoretical guarantees.
- We present a disentangled graph encoder with a tailored training strategy to capture the multiple aspects of the input graph. We further propose an invariance regularized contrastive learning module so that the spurious correlations between the latent factors can be eliminated in the graph representations and achieve representation disentanglement for OOD generalized predictions in downstream tasks.
- We theoretically show that our model can provably learn disentangled graph representations, and make OOD generalization based on the disentangled graph representations.
- Extensive empirical results demonstrate the effectiveness of our proposed **OOD-GCL** against a range of leading-edge baselines on various benchmark datasets under distribution shifts.

The rest of the paper is organized as follows. We introduce some preliminaries in Section 2. Subsequently, in Section 3, we describe the technical details of our proposed **OOD-GCL**. The experimental settings and results are present in Section 4. We review some related works in Section 5. Finally, we conclude this work in Section 6.

2. Preliminaries

Let $\mathbf{G} = \{G_n\}_{n=1}^N$ denote the input graph dataset, where G_n is the n-th graph. The multi-channel graph encoder maps the input graph into its disentangled representation $\mathbf{Z}_n = \Phi(G_n)$. Assuming that there are K latent factors behind the input graph, $\mathbf{Z}_n = [\mathbf{Z}_n^1, \mathbf{Z}_n^2, \dots, \mathbf{Z}_n^K]$ consists of K disentangled components. Let $\mathbf{Z}^k$ denote the representation corresponding to the k-th latent factor for all input graphs. Each $\mathbf{Z}_n^k = \Phi^{(k)}(G_n) \in \mathbb{R}^{\Delta d}$, where $k \in \{1, \dots, K\}$, $\Delta d = d/K$, and d is the representation dimensionality. For simplification, we use $\mathbf{Z}_n^{-k} \in \mathbb{R}^{(K-1)\times \Delta d}$ to denote the concatenated representation except for the representation $\mathbf{Z}_n^k$ corresponding to the k-th latent factor, i.e., $\mathbf{Z}_n^{-k} = [\mathbf{Z}_n^1, \mathbf{Z}_n^2, \dots, \mathbf{Z}_n^{k-1}, \mathbf{Z}_n^{k+1}, \dots, \mathbf{Z}_n^K]$.

Following the invariant learning literature (Li et al., 2022d;b), we make the assumption:

Assumption 1. *There exists a portion of information inside input graph G such that it not only has sufficient abilities in predicting the graph labels, but has invariant relations with the labels under distribution shifts, i.e. satisfying: (a) invariance assumption:* $P^e(Y|\Psi(G)) = P^{e'}(Y|\Psi(G))$

for any environment (domain) e, e' that graph comes from, and $\Psi(\cdot)$ is to capture the invariant information; and (b) sufficiency assumption: $Y = \omega(\Psi(G)) + \epsilon$, where $\omega(\cdot)$ is one function and ϵ is random noise.

Intuitively, if we accurately capture the latent factors that satisfy the invariance assumption and sufficiency assumption, and only make predictions based on these "*invariant & sufficient latent factors*", our predictions can be generalized across different environments under distribution shifts, and the representations corresponding to these latent factors have sufficient abilities in predicting the graph labels. Therefore, the problem can be solved by accurately identifying which latent factor belongs to the "*invariant & sufficient latent factors*" and making OOD generalized predictions only based on the disentangled representations corresponding to these latent factors.

3. Method

In this section, we present the proposed **OOD-GCL** in detail, whose framework is shown in Figure 1. We first present the disentangled graph encoder to identify the complex latent factors and output the factorized representations capturing the multiple aspects of graphs. Then, we propose the disentangled invariant self-supervised learning strategy on tailored contrastive tasks and invariance regularization. Next, we describe how to obtain the predictions based on the learned disentangled representations for specific downstream tasks. Lastly, we provide some theoretical analyses of our proposed method.

3.1. Disentangled Graph Encoder

In this subsection, we present the disentangled graph encoder for producing the factorized graph representation $\mathbf{Z}_n = [\mathbf{Z}_n^1, \mathbf{Z}_n^2, \dots, \mathbf{Z}_n^K]$ for the input graph G_n. After that, we can infer which latent factor belongs to the "*invariant & sufficient latent factors*" and finally make OOD generalized predictions only based on the representations corresponding to these factors.

Typically, graph neural networks (GNNs) leverage the topology of the graph and attributes of nodes to infer the representation vector $\mathbf{h}_v$ for every node v. This is achieved through a message-passing framework, where the representation of a node is recursively refined by aggregating the representations of the adjacent nodes. The message-passing of the l^{th} message-passing layer is formulated as (Li et al., 2021a):

$$\mathbf{h}_v^l = \mathrm{COM}^l\left(\mathbf{h}_v^{l-1}, \mathrm{AGG}^l\left(\{\mathbf{h}_u^{l-1} : u \in \mathcal{N}(v)\}\right)\right), \quad (1)$$

where $\mathbf{h}_v^l$ is the representation of node v at the l^{th} layer and $\mathbf{h}_v^0$ is the input node features. COM and AGG denote the combination and aggregation function, respectively. $\mathcal{N}(v)$ denotes the adjacent nodes of node v. For simplicity, we denote GNN as the message-passing layer in Eq. (1). Define $\mathbf{H}^l$ as the set of node embeddings $\{\mathbf{h}_v^l | v \in V\}$ subsequent to the l^{th} GNN layer, with the set of nodes V in the graph. After the L conventional message-passing layers, our model further adopts a layer for graph disentanglement to output factorized graph representations by extracting features attributed to distinct latent factors through the individual channel. Following (Li et al., 2021a), the complex multiple latent factors are extracted by K different message-passing channels, thereby encapsulating various aspects of the input graph. Specifically, each channel independently employs a GNN_k, using its unique parameters to propagate information: $\mathbf{H}_k^{L+1} = \mathrm{GNN}_k(\mathbf{H}^L, A)$, with A representing the adjacency matrix of the graph. The embeddings $\mathbf{H}_k^{L+1}$ are exclusively associated with the k-th latent factor for disentanglement. Each channel also employs its own READOUT function, essentially a pooling function, to summarize the node embeddings into a graph-level representation: $\mathbf{h}_{G_n,k} = \mathrm{READOUT}_k(\{\mathbf{H}_k^{L+1}\})$. Finally, each channel produces a factor-specific graph representation through an individual MLP: $\mathbf{Z}_n^k = \mathrm{MLP}_k(\mathbf{h}_{G_n,k})$. Following (Li et al., 2021a), the disentangled graph encoder comprises K message-passing channels, which are more specialized in capturing different aspects of the graph, compared with the traditional graph encoders (Kipf & Welling, 2017; Veličković et al., 2018; Xu et al., 2019), that are fundamentally holistic in nature. This design facilitates the identification of intrinsic latent factors and enables the encapsulation of diverse aspects of graphs, which provides a basis for further identifying invariant and sufficient latent factors for OOD generalization.

3.2. Disentangled Invariant Self-supervised Learning

Different from traditional contrastive learning approaches, our proposed **OOD-GCL** introduces an innovative task centered around sufficiently discriminative instance identification of invariant factors. Our model not only draws similar instances closer together while pushing dissimilar ones apart in the representation space to learn sufficiently predictive representations, but it also facilitates the integration of factor-specific information into these representations. This integration enhances the disentanglement of these latent factors, thereby improving the identification of invariant and sufficient latent factors for out-of-distribution (OOD) generalization.

In detail, the formation of real-world graphs is usually driven by multiple latent factors consisting of invariant and sufficient latent factors that can be OOD generalized as well as the other variant latent factors under distribution shifts. Based on the disentangled graph encoder, we can obtain the disentangled graph representations $\mathbf{Z} \in \mathbb{R}^{N \times K \times \Delta d}$ for the whole graph dataset, where $\mathbf{Z}_n^k \in \mathbb{R}^{\Delta d}$ be the k-th chan-

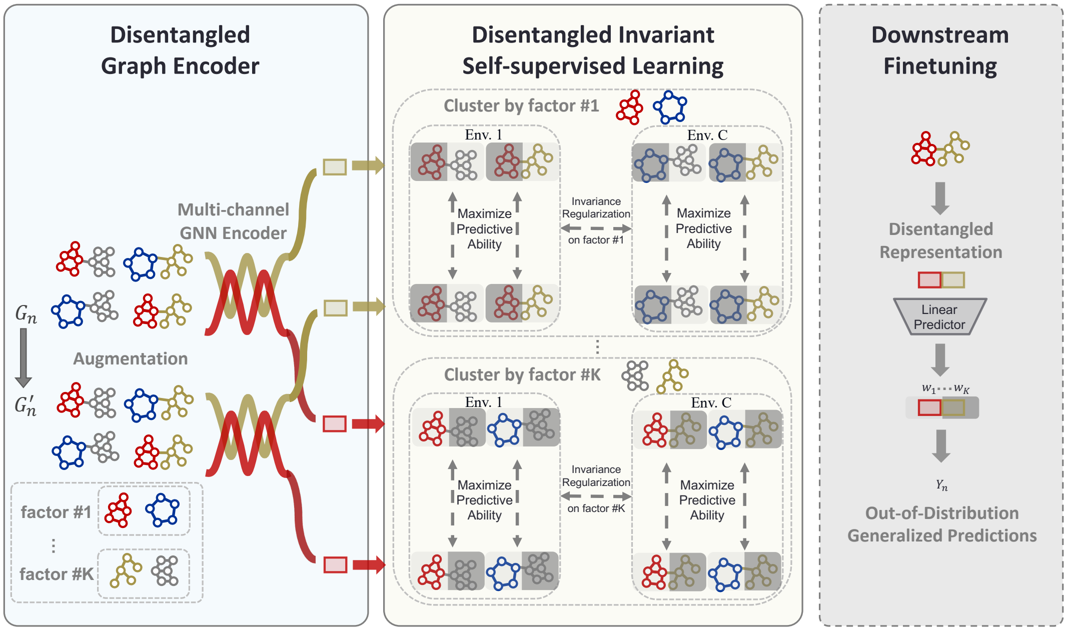


Figure 1. The framework of **OOD-GCL**. It consists of the following stages: (1) The input graph G_n is subjected to graph augmentations, resulting in an augmented view G'_n. Both G_n and G'_n are subsequently processed by a shared multi-channel disentangled graph encoder to produce the factorized graph representations. (2) The disentangled invariant self-supervision learning module is to maximize the predictive ability of the disentangled graph representations and make sure the disentangled graph representations other than that from the k-th channel are invariant under the environment partition inferred by the k-th latent factor, to exclude the information of the k-th latent factor for enhancing representation disentanglement. (3) At the downstream finetuning stage, our model can achieve theoretically guaranteed OOD generalization ability with a linear predictor based on the learned disentangled graph representations.

nel's disentangled representation for G_n. We first infer the environment partition $\mathcal{E}^k$ by clustering the representation corresponding to the k-th latent factor for all graphs $\mathbf{Z}^k$ as follows:

$$\mathcal{E}^k = \text{Cluster}(\mathbf{Z}^k). \tag{2}$$

Here the motivation is that since the disentangled graph representation of each channel only captures the specific information in terms of each latent factor, the representations should naturally form multiple clusters (i.e., partitions) according to their intrinsic semantic information of the latent factor.

Specifically, we adopt the k-means (MacQueen et al., 1967) clustering algorithm by setting the cluster number C in an adaptive manner. The Eq. (2) is implemented by:

$$C^* = \text{argmax}_C \ \text{Sil}(\mathcal{E}^k), \tag{3}$$

where $\text{Sil}(\cdot)$ is the Silhouette Score (Rousseeuw, 1987) that is a commonly used metric for evaluating the quality of clustering. It measures how similar each sample is to its own cluster compared to other clusters. The score ranges from -1 to 1, where a higher score indicates better-defined clusters. By calculating this Silhouette Score for different numbers of clusters, we can adaptively determine the optimal number of clusters by maximizing the score, thus ensuring the most appropriate cluster configuration. Finally, we obtain the environment partition with the optimal number of clustering, i.e.,

$$\mathcal{E}^k = \text{k-means}(\mathbf{Z}^k; C^*). \tag{4}$$

We expect the inferred environment partition can reflect the intrinsic clusters in semantic space of the latent factor that could be the label information in the downstream tasks (Wang et al., 2021a).

After obtaining the inferred environment partition, we utilize the disentangled representation $\mathbf{Z}_n^{-k}$ that is from all channels except for the k-th channel to calculate the contrastive loss which can be regarded as discriminating oneself from several negative samples. We define the loss function ℓ_e^k in the e-th environment (i.e., cluster) inferred by the disentangled representation of the k-th latent factor as follows:

$$\ell_e^k = -\sum_{n=1}^{N} \mathbb{1}_{[\psi_n = e]} \log \frac{\exp \phi(\mathbf{Z}_n^{-k}, \mathbf{Z}_n'^{-k})}{\sum_{n' \neq n} \mathbb{1}_{[\psi_{n'} = e]} \exp \phi(\mathbf{Z}_n^{-k}, \mathbf{Z}_{n'}'^{-k})}, \tag{5}$$

where $e \in \{1, \ldots, |\mathcal{E}^k|\}$, $|\mathcal{E}^k|$ denotes the number of clusters, and $n' \in \{1, \ldots, N\}$. $\mathbb{1}_{[\psi_n = e]} \in \{0, 1\}$ is an indicator function evaluating to 1 *iff.* $\psi_n = e$ and ψ_n is the environment of the n-th graph considering the current environment partition. ϕ is the cosine similarity with temperature τ, i.e., $\phi(\mathbf{a}, \mathbf{b}) = \cos(\mathbf{a}, \mathbf{b})/\tau$ and $\cos(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b} / (\|\mathbf{a}\|_2 \|\mathbf{b}\|_2)$. $\mathbf{Z}_n'^{-k}$ is the corresponding representation from the augmented view of the n-th graph except for the k-th channel. We adopt the common augmentation strategies following (You et al., 2020) to obtain $\mathbf{Z}_n'^{-k}$. Overall, the training objective is formulated as:

$$\ell = \sum_{k=1}^{K} \ell^k, \tag{6}$$

where

$$\ell^k = \sum_{e \in \mathcal{E}^k} \ell_e^k + \lambda \mathrm{Inv}(\ell_e^k; \mathcal{E}^k). \tag{7}$$

$\mathrm{Inv}(\cdot)$ denotes the invariance regularization. In our method, we adopt V-REx (Krueger et al., 2021) as our invariance regularization and λ is the invariance coefficient. Note that our method is also compatible with the other invariance regularization, e.g., IRM (Arjovsky et al., 2019), IGA (Koyama & Yamaguchi, 2020), etc.

Intuitively, the first contrastive loss term in Eq. (7) is to encourage sufficient predictive ability of the disentangled graph representations, while the second invariance regularization term in Eq. (7) is to make sure the disentangled graph representations other than that from the k-th channel are invariant under the environment partition inferred by the k-th latent factor, so as to exclude the information of the k-th latent factor for representation disentanglement.

Finally, we iteratively update $\mathcal{E}^k$ by Eq. (4) and optimize the disentangled graph encoder by Eq. (6) until convergence.

3.3. Downstream Finetuning

Based on the disentangled invariant self-supervised learning, the graph encoder can be optimized to make the representation disentangled in a self-supervised manner. After that, when deployed in the specific downstream tasks, the disentangled graph representations are further fed into a linear predictor ω to calculate the graph labels. We finetune the disentangled graph representations and linear predictor to make predictions to the graph labels as:

$$\hat{Y}_n = \omega(\mathbf{Z}_n). \tag{8}$$

Note that although more powerful predictors (e.g., MLP, etc.) can also be utilized to map the disentangled graph representations into the labels, we show that such a simple linear predictor is good enough to achieve OOD generalization in the next subsection.

3.4. Theoretical Analyses

Here we provide some theoretical analyses from two aspects, which show: (1) our model can provably learn disentangled graph representations, and (2) our model can make OOD generalization based on the output disentangled graph representations.

First, we analyze why our model can learn disentangled graph representations. Note that in this work, we follow (Wang et al., 2021a) to adopt the established definition of disentanglement (Higgins et al., 2018).

Definition 1. (Disentangled Representation (Higgins et al., 2018; Wang et al., 2021a)) *Let $\mathcal{U}$ denote the semantic space. The semantic information potentially corresponds to the label of the downstream task. Let $\mathcal{Z}$ denote the representation space. Let $\mathcal{G}$ denote the group that acts on $\mathcal{U}$. Assume that there is a direct product decomposition where $\mathcal{G}$ is decomposed into $\mathcal{G}^1 \times \ldots \times \mathcal{G}^K$, and $\mathcal{U}$ is decomposed into $\mathcal{U}^1 \times \ldots \times \mathcal{U}^K$, with $\mathcal{G}^k$ acting on $\mathcal{U}^k$ respectively. The representation is disentangled if it satisfies the following equivariant property and decomposable property in the meantime.*

1. *Equivariant property*: $\forall g \in \mathcal{G}, \forall u \in \mathcal{U}, f(g \cdot u) = g \cdot f(u)$, *where $f : \mathcal{U} \to \mathcal{Z}$ is the representation function.*
2. *Decomposable property*: *there is a decomposition $\mathcal{Z} = \mathcal{Z}^1 \times \ldots \times \mathcal{Z}^K$, where each $\mathcal{Z}^k$ is affected only by $\mathcal{G}^k$ and unaffected by all $\mathcal{G}^{k'}$, $k \neq k'$, and k, $k' \in \{1, \ldots, K\}$.*

Intuitively, the equivariant property means that the action of $\mathcal{G}$ on semantic space $\mathcal{U}$ is equivariant to the action on $\mathcal{Z}$. The decomposable property means that the representation corresponding to one latent factor is not affected by changing the semantic information of the other latent factors.

Let g denote a group element in $\mathcal{G}$ and g can be decomposed by $(g^1, g^2, ..., g^K)$, where $g^k \in \mathcal{G}^k$. The goal of our method is to learn the disentangled graph representations *w.r.t.* $\prod_{k=1}^{K} \mathcal{G}^k$. The group action $g \in \mathcal{G}$ on the semantic space $\mathcal{U}$ is equivariant to its action on the representation $\mathcal{Z}$. Furthermore, the representation space $\mathcal{Z}$ is decomposed into K parts: $\mathcal{Z}^1 \times \ldots \times \mathcal{Z}^K$, where $\mathcal{Z}^k$ is affected only by $g^k \in \mathcal{G}^k$ and unaffected by all $g^{k'} \in \mathcal{G}^{k'} (k' \neq k)$. Inspired by the literature (Wang et al., 2021a), in the following theorem, we prove that optimizing the objective function of our method can lead to disentangled graph representations, formulated by this definition.

Theorem 1. *The graph representation is disentangled w.r.t. $\prod_{k=1}^{K} \mathcal{G}^k$ if the objective function Eq.* (7) *reaches the minimum for any k-th latent factor, $k \in \{1, \ldots, K\}$.*

The proof is shown in Appendix due to the page limit.

Next, we prove that the disentangled graph representations can lead to OOD generalized predictions under mild assumptions. Following the literature (Xu et al., 2022), we want to estimate the coefficient of the linear predictor to calculate the label given the disentangled graph representations.

Theorem 2. *Given the disentangled graph representations, (1) if the k-th latent factor does not belong to the invariant and sufficient latent factors to predict the label Y, then the coefficient of the linear predictor w^k equals zero; and (2) if the k-th latent factor belongs to the invariant and sufficient latent factors to predict the label Y, then the coefficient of the linear predictor w^k does not equal zero.*

The proof is also shown in Appendix. Intuitively, for achieving OOD generalization, we aim to identify the invariant and sufficient latent factors under distribution shifts, so that we can learn a graph model that is generalized to unknown testing distribution. As relationships between entangled graph representations usually change under distribution shifts, the model will fail to achieve OOD generalization if we use traditional entangled representations. Therefore, it is critical to find the invariant and sufficient latent factors and adopt their disentangled representations for making predictions, so that it can relieve the negative impact from the other latent factors that are not in the invariant and sufficient latent factors under distribution shifts.

4. Experiments

In this section, we conduct experiments on real-world datasets to show the effectiveness of the proposed **OOD-GCL** model for handling distribution shifts on graphs.

4.1. Experimental Setup

Datasets. We adopt real-world benchmark datasets for the graph classification task, including the datasets from graph OOD generalization benchmark GOOD (Gui et al., 2022) and the datasets from Open Graph benchmark (Hu et al., 2020). Following (Sui et al., 2023), the datasets in the experiment consist of:

- Motif. Each graph in this dataset consists of a motif and a base graph, where the ground-truth label only depends on the motif. The distribution shift is induced by different spurious correlations between the label and the base graph, as well as different sizes.
- CMNIST. Each graph is converted from an image in MNIST (LeCun et al., 1998) which is to be classified into the corresponding handwritten digit. The distribution shift exists on node features by colorizing the digits differently.
- Molbbbp and Molhiv. The distribution shift exists on graph structural properties (scaffolds) or sizes. The provided split separates structurally different molecules with different scaffolds or sizes into different subsets.

Note that all of the datasets consist of two data split strategies to create different distribution shifts except for CMNIST, following the well-established settings (Gui et al., 2022; Sui et al., 2023).

Baselines. We compare our **OOD-GCL** with several representative state-of-the-art methods, which can be divided into three groups:

- General invariant learning methods, including standard ERM, IRM (Arjovsky et al., 2019), GroupDRO (Sagawa et al., 2019), V-REx (Krueger et al., 2021).
- Graph OOD generalization methods, including DIR (Wu et al., 2022b), CAL (Sui et al., 2022), GSAT (Miao et al., 2022), OOD-GNN (Li et al., 2022a), StableGNN (Fan et al., 2021), CIGA (Chen et al., 2022), DisC (Fan et al., 2022), DropEdge (Rong et al., 2019), GREA (Liu et al., 2022), FLAG (Kong et al., 2022), M-Mixup (Wang et al., 2021b), $\mathcal{G}$-Mixup (Han et al., 2022), AIA (Sui et al., 2023).
- Graph self-supervised learning methods, including InfoGraph (Sun et al., 2019), GraphCL (You et al., 2020), GMI (Peng et al., 2020), CNC (Zhang et al., 2022a).

Implementation Details. The number of epochs for pretraining and finetuning is chosen from $\{50, 100, 200\}$. The Adam optimizer (Kingma & Ba, 2014) is adopted for gradient descent. The evaluation metric is accuracy for Motif and CMNIST datasets and ROC-AUC for Molbbbp and Molhiv datasets. The dimensionality of the representations d is chosen from $[128, 256, 512]$. The invariance regularizer coefficient λ is chosen from $\{10^{-4}, 10^{-2}, 10^{0}\}$. The number of disentangled channels K is chosen from $\{2, 3, 4, 5\}$. We report mean results and standard deviations of ten runs.

4.2. Main Results

The comparisons of different methods on the real-world graph datasets under distribution shifts are shown in Table 1. We have the following observations.

- Graph OOD generalization methods achieve better performance than general invariant learning methods, e.g., IRM, V-REx, etc. Such performance gains demonstrate the importance of considering and addressing the complex structural distribution shifts on graphs in the model design.
- Graph self-supervised learning methods, despite lacking tailored designs for handling graph distribution shifts, exhibit comparable performance to the graph OOD generalization methods in some comparisons. This suggests that the inherent capacity of graph self-supervised learning to capture truly predictive structural information in graph data contributes to its effectiveness in tackling distribution shifts.

Table 1. Comparisons of different methods on the real-world graph datasets under distribution shifts. Numbers after the ± signs represent the standard deviations. The best results are in bold. The results show that our method outperforms the baseline methods consistently, including general invariant learning methods, graph OOD generalization methods, and graph self-supervised learning methods.

Dataset	Motif		CMNIST	Molbbbp		Molhiv	
	base	size	color	scaffold	size	scaffold	size
ERM	68.66±4.25	51.74±2.88	28.60±1.87	68.10±1.68	78.29±3.76	69.58±2.51	59.94±2.37
IRM	70.65±4.17	51.41±3.78	27.83±2.13	67.22±1.15	77.56±2.48	67.97±1.84	59.00±2.92
GroupDRO	68.24±8.92	51.95±5.86	29.07±3.14	66.47±2.39	79.27±2.43	70.64±2.57	58.98±2.16
V-REx	71.47±6.69	52.67±5.54	28.48±2.87	68.74±1.03	78.76±2.37	70.77±2.84	58.53±2.88
DIR	62.07±8.75	52.27±4.56	33.20±6.17	66.86±2.25	76.40±4.43	68.07±2.29	58.08±2.31
CAL	65.63±4.29	51.18±5.60	27.99±3.24	68.06±2.60	79.50±4.81	67.37±3.61	57.95±2.24
GSAT	62.80±11.41	53.20±8.35	28.17±1.26	66.78±1.45	75.63±3.83	68.66±1.35	58.06±1.98
OOD-GNN	61.10±7.87	52.61±4.67	26.49±2.94	66.72±1.23	79.48±4.19	70.46±1.97	60.60±3.77
StableGNN	57.07±14.10	46.93±8.85	28.38±3.49	66.74±1.30	77.47±4.69	68.44±1.33	56.71±2.79
CIGA	66.43±11.31	49.14±8.34	32.22±2.67	64.92±2.09	65.98±3.31	69.40±2.39	59.55±2.56
DisC	51.08±3.08	50.39±1.15	24.99±1.78	67.12±2.11	56.59±10.09	68.07±1.75	58.76±0.91
DropEdge	45.08±4.46	45.63±4.61	22.65±2.90	66.49±1.55	78.32±3.44	70.78±1.38	58.53±1.26
GREA	56.74±9.23	54.13±10.02	29.02±3.26	69.72±1.66	77.34±3.52	67.79±2.56	60.71±2.20
FLAG	61.12±5.39	51.66±4.14	32.30±2.69	67.69±2.36	79.26±2.26	68.45±2.30	60.59±2.95
M-Mixup	70.08±3.82	51.48±4.91	26.47±3.45	68.75±0.34	78.92±2.43	68.88±2.63	59.03±3.11
$\mathcal{G}$-Mixup	59.66±7.03	52.81±6.73	31.85±5.82	67.44±1.62	78.55±4.16	70.01±2.52	59.34±2.43
AIA	73.64±5.15	55.85±7.98	36.37±4.44	70.79±1.53	81.03±5.15	71.15±1.81	61.64±3.37
InfoGraph	69.73±2.74	50.17±3.72	33.84±1.52	70.39±1.34	80.82±0.49	67.51±3.65	60.16±2.12
GraphCL	71.34±3.76	52.77±5.89	32.81±1.71	69.36±1.32	80.64±0.78	70.12±2.19	57.19±3.18
GMI	65.26±1.98	51.85±3.49	30.24±5.98	69.38±1.02	77.67±0.30	68.95±2.92	59.86±2.78
CNC	70.34±2.73	50.64±4.86	32.41±1.28	68.16±1.25	76.19±3.52	69.13±3.71	58.67±3.41
OOD-GCL	**76.31±2.37**	**60.12±2.35**	**41.89±2.03**	**72.86±1.97**	**83.12±3.21**	**74.13±2.01**	**63.41±2.59**

- Our method outperforms the baseline methods consistently, including general invariant learning methods, graph OOD generalization methods, and graph self-supervised learning methods. This superiority can be attributed to the ability to effectively identify underlying invariant and sufficient latent factors that are crucial for preserving truly predictive graph properties and learning disentangled representations. Unlike the baseline methods, our approach employs disentangled invariant self-supervised learning during the pretraining stage, leading that the graph representations can be disentangled and OOD generalized. This unique design enables our method to more effectively generalize to unknown distribution shifts, ultimately resulting in superior performance across a range of graph prediction tasks under distribution shifts. Overall, the results highlight the significance of learning disentangled representations in addressing challenges posed by graph distribution shifts.

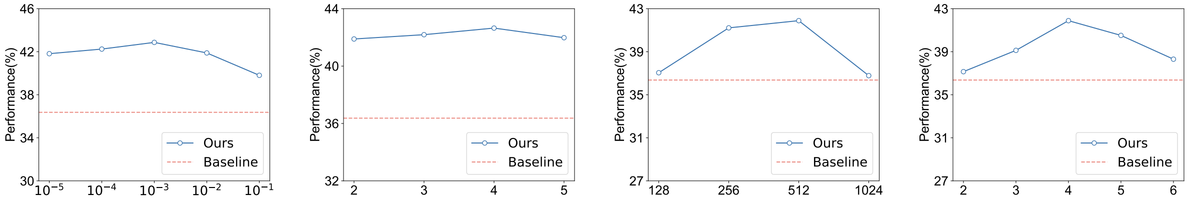
4.3. Time Complexity Analysis

Now we analyze the time complexity of our method to show its efficiency. The time complexity of the proposed **OOD-GCL** is $O(|E|\, d + |V|\, d^2)$, where $|V|$ represents the number of nodes, $|E|$ denotes the number of edges, and d is the dimensionality of the representations. In particular, we employ the message-passing GNN to instantiate the GNN components in the disentangled graph encoder, which also exhibits a complexity of $O(|E|\, d + |V|\, d^2)$. As for the disentangled invariant self-supervised learning module, the time complexity is $O(|\mathcal{B}|^2 d)$, where $|\mathcal{B}|$ is the batch size. Since $|\mathcal{B}|$ is small constant, the overall time complexity of **OOD-GCL** is $O(|E|\, d + |V|\, d^2)$. In comparison, the time complexity of the other GNN-based baselines is also $O(|E|\, d + |V|\, d^2)$. Therefore, the time complexity of our **OOD-GCL** is comparable with the existing baselines. Although our method exhibits effectiveness, it does not induce a higher time cost.

4.4. Ablation Study

In this subsection, we conduct ablation studies to verify the effectiveness of the key modules in our proposed **OOD-GCL**. We consider the following ablated variants:

- Variant “w/o Inv.” means that we set $\lambda = 0$ in Eq. (7),



(a) Invariance coefficient λ (b) Number of factors K (c) Dimensionality d (d) Number of layers L

Figure 2. Hyperparameter sensitivity analysis on CMNIST dataset. The blue lines denote the results of our method and the red dashed lines are the results of the best baseline.

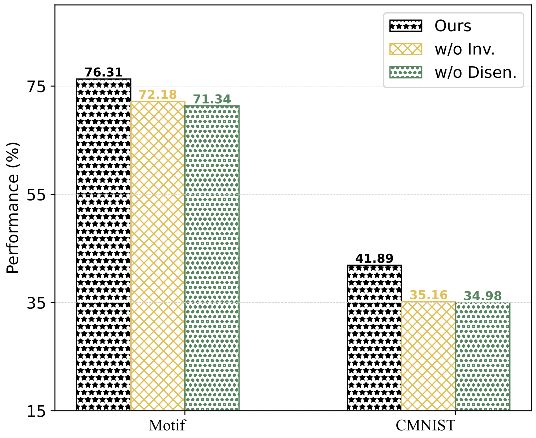


Figure 3. Ablation studies. The ablated version "w/o Inv." removes the invariance regularization by setting $\lambda = 0$ in Eq. (7), and "w/o Disen." removes the disentangled module and degenerates to the entangled version.

namely removing the invariance regularization.

- Variant "w/o Disen." means that we remove the disentanglement design and the model is degenerated to the entangled version.

The results of **OOD-GCL** and its variants are shown in Figure 3. Notably, there is a noticeable performance drop for the variant "w/o Inv.", indicating the effectiveness of inferring and identifying the latent factor behind graphs for each channel accurately. The invariance regularization term ensures that the k-th part of disentangled graph representations only includes the information of the k-th latent factor. So it will affect the disentanglement if this invariance regularization is removed. In the case of "w/o Disen.", the absence of disentanglement leads to entangled latent factors within the graph representations. This entanglement poses challenges in characterizing distinct aspects of the graphs and hinders the ability to learn informative graph representations and achieve OOD generalization.

4.5. Hyperparameter Sensitivity

In this subsection, we analyze the sensitivity of some important hyperparameters. First, the hyperparameter λ in Eq. (7) has a moderate influence on the performance, since λ serves a role in encouraging the model to capture the invariance among different environments. To explore the sensitivity of our method on this hyperparameter, we adjust $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, while maintaining the default value of the other hyperparameters unchanged. The results on the CMNIST dataset are shown in Figure 2, while the results on the other datasets show similar patterns. The results show that a small λ may struggle to capture disentangled latent factors to output representations accurately, leading to challenges in achieving OOD generalized predictions. Conversely, a large λ can lead that the learned representations are not informative enough to capture the graph properties. Besides, we also conduct sensitivity analysis on the other hyper-parameters, including the number of factors $K \in \{2, 3, 4, 5\}$, the dimensionality $d \in \{128, 256, 512, 1024\}$, and the number of GNN layers $L \in \{2, 3, 4, 5, 6\}$. The results show that our method outperforms the best baseline in a wide range of hyperparameters, showing our method is not sensitive to hyperparameter choices.

5. Related Work

OOD Generalization on Graphs. A basic assumption for most machine learning methods is that there exists an identical distribution between training and testing data. However, this assumption is frequently violated in real-world scenarios due to inevitable distribution shifts, thereby introducing substantial challenges to the model's generalization performance, particularly in out-of-distribution (OOD) scenarios (Shen et al., 2021). The model performance decline becomes noticeable when it lacks enough OOD generalization capability. To tackle this problem, several general invariant learning methods have been developed (Arjovsky et al., 2019; Krueger et al., 2021; Koyama & Yamaguchi, 2020; Chang et al., 2020; Creager et al., 2021; Liu et al., 2021a). Recently, OOD generalization on graphs (Li et al., 2022b) has also attracted much attention, considering the complex distribution shifts existing on the graph structure and node feature. Several famous graph OOD generalization work (Wu et al., 2022c;a; Li et al., 2022d; Fan et al., 2022; Sui et al., 2022; Liu et al., 2022; Sui et al., 2023; Zhang et al., 2022b; 2023b;c;d; Li et al., 2023; Mao et al., 2024) have been proposed, in general focusing on learning invariant graph patterns among different environments for tackling the distribution shifts. Most of the approaches typically adopt a supervised manner to get rid of the spurious correlations between the variant patterns and the labels and facilitate the invariant patterns for predictions. However, they heavily rely on sufficient task-dependent annotated labels to learn OOD generalized graph representations, which could be unavailable in practice. To this end, in this paper, we focus on the self-supervised graph OOD generalization problem, namely learning GNNs capable of achieving relatively stable performances under distribution shifts without graph labels. This problem is not explored in the literature.

Contrastive Learning on Graphs. Self-supervised learning is one machine learning technique to learn informative representations by training the model using the data itself rather than the external annotated labels (Liu et al., 2021b). Recently, self-supervised learning has been revolutionized by the development of contrastive learning, which primarily focuses on instance discrimination as a pretext task (Jaiswal et al., 2020). The methodologies on this technique are widely developed (Oord et al., 2018; Wu et al., 2018; He et al., 2020; Chen et al., 2020; Tsai et al., 2021). In the meantime, the application of contrastive learning on graph data has also been explored in various research works (Sun et al., 2019; Qiu et al., 2020; Hassani & Khasahmadi, 2020; You et al., 2020; Zhang et al., 2021; Li et al., 2021a; Ren & Liu, 2020; Peng et al., 2020; Li et al., 2022e;c). The fundamental principle of these approaches is to enhance the similarity or agreement between different views of the input graph. Nevertheless, current methods in graph contrastive learning primarily focus on general settings and ignore the diverse latent factors within the complex graph data. Consequently, such approaches may not adequately preserve intrinsic and truly predictive graph characteristics, leading to the potential fragility to distribution shifts.

Disentangled Representation Learning. Disentangled representation learning is expected to capture the latent factors behind the observed data, where these latent factors are represented as factorized vectors (Bengio et al., 2013; Higgins et al., 2018; Wang et al., 2022a; Pfau et al., 2020; Winter et al., 2022; Shu et al., 2018). It has found widespread applications in various domains, including computer vision (Hsieh et al., 2018; Wang et al., 2021a; 2023b; Chen et al., 2023), recommendation systems (Wang et al., 2022b; 2023a; Zhang et al., 2023a; Wang et al., 2024), etc. Noteworthy advancements in this field also involve the application of disentangled representation learning for graph data through the development of disentangled graph neural networks (Ma et al., 2019; Li et al., 2021b; 2022c; Yang et al., 2020; Zhang et al., 2023e; 2024b;a). For instance, (Ma et al., 2019) introduce a neighborhood routing mechanism within graph convolution to identify latent factors influencing node-to-neighbor connections. (Liu et al., 2020) strive to enhance the independence of graph latent factors by eliminating their mutual dependence, leveraging a kernel-based metric. Some works (Li et al., 2021b; 2022c; Wang et al., 2021a) extend the disentangled representation learning in a self-supervised manner, but they fail to tackle the distribution shifts for OOD generalization with a theoretical guarantee. We draw inspirations from the concept of disentangled representation learning (Higgins et al., 2018; Wang et al., 2021a; 2022a) to address the graph out-of-distribution generalization problems and bridge the disentanglement and OOD generalization on graphs from a theoretical view.

6. Conclusion

In this paper, we study the self-supervised graph OOD generalization problem. The key challenge lies in disentangling invariant and variant latent factors behind the graphs without graph labels for tackling distribution shifts. To address the challenge, we propose a novel OOD generalized disentangled graph contrastive learning model (**OOD-GCL**), which is able to learn disentangled graph representations with self-supervision that can generalize under distribution shifts. We conduct comprehensive theoretical analyses to prove that our method can learn disentangled graph representations and achieve OOD generalization. The extensive experiments on several real-world benchmarks demonstrate the superiority of the proposed method over state-of-the-art baselines.

Acknowledgements

This work is supported by the National Key Research and Development Program of China No.2023YFF1205001, National Natural Science Foundation of China (No. 62222209, 62250008, 62102222, 62206149), Beijing National Research Center for Information Science and Technology under Grant No. BNR2023RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *TPAMI*, 35(8): 1798–1828, 2013.

Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. Invariant rationalization. In *International Conference on Machine Learning*, pp. 1448–1458. PMLR, 2020.

Chen, H., Zhang, Y., Wu, S., Wang, X., Duan, X., Zhou, Y., and Zhu, W. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607. PMLR, 2020.

Chen, Y., Zhang, Y., Bian, Y., Yang, H., Ma, K., Xie, B., Liu, T., Han, B., and Cheng, J. Learning causally invariant representations for out-of-distribution generalization on graphs. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Fan, S., Wang, X., Shi, C., Cui, P., and Wang, B. Generalizing graph neural networks on out-of-distribution graphs. *arXiv preprint arXiv:2111.10657*, 2021.

Fan, S., Wang, X., Mo, Y., Shi, C., and Tang, J. Debiasing graph neural networks via learning disentangled causal substructure. *Advances in Neural Information Processing Systems*, 35:24934–24946, 2022.

Gui, S., Li, X., Wang, L., and Ji, S. Good: A graph out-of-distribution benchmark. *Advances in Neural Information Processing Systems*, 35:2059–2073, 2022.

Han, X., Jiang, Z., Liu, N., and Hu, X. G-mixup: Graph data augmentation for graph classification. In *International Conference on Machine Learning*, pp. 8230–8248. PMLR, 2022.

Hassani, K. and Khasahmadi, A. H. Contrastive multi-view representation learning on graphs. In *ICML*, pp. 4116–4126. PMLR, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv:1812.02230*, 2018.

Hsieh, J.-T., Liu, B., Huang, D.-A., Fei-Fei, L. F., and Niebles, J. C. Learning to decompose and disentangle representations for video prediction. In *NeurIPS*, pp. 517–526, 2018.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. *Neural Information Processing Systems (NeurIPS)*, 2020.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Kong, K., Li, G., Ding, M., Wu, Z., Zhu, C., Ghanem, B., Taylor, G., and Goldstein, T. Robust optimization as data augmentation for large-scale graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 60–69, 2022.

Koyama, M. and Yamaguchi, S. Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint arXiv:2008.01883*, 2020.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, H., Wang, X., Zhang, Z., Yuan, Z., Li, H., and Zhu, W. Disentangled contrastive learning on graphs. In *NeurIPS*, 2021a.

Li, H., Wang, X., Zhang, Z., Yuan, Z., Li, H., and Zhu, W. Disentangled contrastive learning on graphs. *Advances in Neural Information Processing Systems*, 34:21872–21884, 2021b.

Li, H., Wang, X., Zhang, Z., and Zhu, W. Ood-gnn: Out-of-distribution generalized graph neural network. *IEEE Transactions on Knowledge and Data Engineering*, 2022a.

Li, H., Wang, X., Zhang, Z., and Zhu, W. Out-of-distribution generalization on graphs: A survey. *arXiv preprint arXiv:2202.07987*, 2022b.

Li, H., Zhang, Z., Wang, X., and Zhu, W. Disentangled graph contrastive learning with independence promotion. *IEEE Transactions on Knowledge and Data Engineering*, 2022c.

Li, H., Zhang, Z., Wang, X., and Zhu, W. Learning invariant graph representations for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:11828–11841, 2022d.

Li, H., Zhang, Z., Wang, X., and Zhu, W. Invariant node representation learning under distribution shifts with multiple latent environments. *ACM Transactions on Information Systems*, 42(1):1–30, 2023.

Li, S., Wang, X., Zhang, A., Wu, Y., He, X., and Chua, T.-S. Let invariant rationale discovery inspire graph contrastive learning. In *International conference on machine learning*, pp. 13052–13065. PMLR, 2022e.

Liu, G., Zhao, T., Xu, J., Luo, T., and Jiang, M. Graph rationalization with environment-based augmentations. In *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2022.

Liu, J., Hu, Z., Cui, P., Li, B., and Shen, Z. Heterogeneous risk minimization. In *International Conference on Machine Learning*. PMLR, 2021a.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021b.

Liu, Y., Wang, X., Wu, S., and Xiao, Z. Independence promoted graph disentangled networks. In *AAAI*, 2020.

Ma, J., Cui, P., Kuang, K., Wang, X., and Zhu, W. Disentangled graph convolutional networks. In *International conference on machine learning*, pp. 4212–4221. PMLR, 2019.

MacQueen, J. et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 281–297. Oakland, CA, USA, 1967.

Mao, W., Wu, J., and Wang, X. InfoIGL: Invariant graph learning driven by information theory. 2024. URL https://openreview.net/forum?id=vOOkWxbLs7.

Miao, S., Liu, M., and Li, P. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*, pp. 15524–15543, 2022.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., and Tekade, R. K. Artificial intelligence in drug discovery and development. *Drug discovery today*, 26(1):80, 2021.

Peng, Z., Huang, W., Luo, M., Zheng, Q., Rong, Y., Xu, T., and Huang, J. Graph representation learning via graphical mutual information maximization. In *WebConf*, pp. 259–270, 2020.

Pfau, D., Higgins, I., Botev, A., and Racanière, S. Disentangling by subspace diffusion. *Advances in Neural Information Processing Systems*, 33:17403–17415, 2020.

Qiu, J., Tang, J., Ma, H., Dong, Y., Wang, K., and Tang, J. Deepinf: Social influence prediction with deep learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2110–2119, 2018.

Qiu, J., Chen, Q., Dong, Y., Zhang, J., Yang, H., Ding, M., Wang, K., and Tang, J. Gcc: Graph contrastive coding for graph neural network pre-training. In *SIGKDD*, pp. 1150–1160, 2020.

Ren, Y. and Liu, B. Heterogeneous deep graph infomax. In *AAAI Workshop of Deep Learning on Graphs*, 2020.

Rong, Y., Huang, W., Xu, T., and Huang, J. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. Towards out-of-distribution generalization: A survey. *arXiv:2108.13624*, 2021.

Shu, Z., Sahasrabudhe, M., Guler, R. A., Samaras, D., Paragios, N., and Kokkinos, I. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 650–665, 2018.

Sui, Y., Wang, X., Wu, J., Lin, M., He, X., and Chua, T.-S. Causal attention for interpretable and generalizable graph classification. *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2022.

Sui, Y., Wu, Q., Wu, J., Cui, Q., Li, L., Zhou, J., Wang, X., and He, X. Unleashing the power of graph data augmentation on covariate distribution shift. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Sun, F.-Y., Hoffman, J., Verma, V., and Tang, J. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*, 2019.

Tsai, T. W., Li, C., and Zhu, J. Mi{ce}: Mixture of contrastive experts for unsupervised image clustering. In *ICLR*, 2021.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Wang, T., Yue, Z., Huang, J., Sun, Q., and Zhang, H. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34:18225–18240, 2021a.

Wang, X., Chen, H., Tang, S., Wu, Z., and Zhu, W. Disentangled representation learning. *arXiv preprint arXiv:2211.11695*, 2022a.

Wang, X., Pan, Z., Zhou, Y., Chen, H., Ge, C., and Zhu, W. Curriculum co-disentangled representation learning across multiple environments for social recommendation. In *International Conference on Machine Learning*, pp. 36174–36192. PMLR, 2023a.

Wang, X., Wu, Z., Chen, H., Lan, X., and Zhu, W. Mixup-augmented temporally debiased video grounding with content-location disentanglement. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4450–4459, 2023b.

Wang, Y., Wang, W., Liang, Y., Cai, Y., and Hooi, B. Mixup for node and graph classification. In *Proceedings of the Web Conference 2021*, pp. 3663–3674, 2021b.

Wang, Y., Qin, Y., Sun, F., Zhang, B., Hou, X., Hu, K., Cheng, J., Lei, J., and Zhang, M. Disenctr: Dynamic graph-based disentangled representation for click-through rate prediction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2314–2318, 2022b.

Wang, Y., Wang, X., Huang, X., Yu, Y., Li, H., Zhang, M., Guo, Z., and Wu, W. Intent-aware recommendation via disentangled graph contrastive learning. *arXiv preprint arXiv:2403.03714*, 2024.

Winter, R., Bertolini, M., Le, T., Noe, F., and Clevert, D.-A. Unsupervised learning of group invariant and equivariant representations. *Advances in Neural Information Processing Systems*, 35:31942–31956, 2022.

Wu, Q., Zhang, H., Yan, J., and Wipf, D. Handling distribution shifts on graphs: An invariance perspective. *International Conference on Learning Representations*, 2022a.

Wu, Y., Wang, X., Zhang, A., He, X., and Chua, T.-S. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations*, 2022b.

Wu, Y.-X., Wang, X., Zhang, A., He, X., and seng Chua, T. Discovering invariant rationales for graph neural networks. In *ICLR*, 2022c.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pp. 3733–3742, 2018.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Xu, R., Zhang, X., Shen, Z., Zhang, T., and Cui, P. A theoretical analysis on independence-driven importance weighting for covariate-shift generalization. In *International Conference on Machine Learning*, pp. 24803–24829. PMLR, 2022.

Yang, S., Zhang, Z., Zhou, J., Wang, Y., Sun, W., Zhong, X., Fang, Y., Yu, Q., and Qi, Y. Financial risk analysis for smes with graph-based supply chain mining. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 4661–4667, 2021.

Yang, Y., Feng, Z., Song, M., and Wang, X. Factorizable graph convolutional networks. In *NeurIPS*, 2020.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., and Shen, Y. Graph contrastive learning with augmentations. In *NeurIPS*, 2020.

Yu, B., Yin, H., and Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022a.

Zhang, S., Hu, Z., Subramonian, A., and Sun, Y. Motif-driven contrastive learning of graph representations. *arXiv:2012.12533*, 2021.

Zhang, Y., Wang, X., Chen, H., and Zhu, W. Adaptive disentangled transformer for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3434–3445, 2023a.

Zhang, Z., Wang, X., Zhang, Z., Li, H., Qin, Z., and Zhu, W. Dynamic graph neural networks under spatio-temporal distribution shift. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022b.

Zhang, Z., Li, X., Teng, F., Lin, N., Zhu, X., Wang, X., and Zhu, W. Out-of-distribution generalized dynamic graph neural network for human albumin prediction. In *IEEE International Conference on Medical Artificial Intelligence*, 2023b.

Zhang, Z., Wang, X., Zhang, Z., Li, H., and Zhu, W. Out-of-distribution generalized dynamic graph neural network with disentangled intervention and invariance promotion. *arXiv preprint arXiv:2311.14255*, 2023c.

Zhang, Z., Wang, X., Zhang, Z., Qin, Z., Wen, W., Xue, H., Li, H., and Zhu, W. Spectral invariant learning for dynamic graphs under distribution shifts. In *Advances in Neural Information Processing Systems*, 2023d.

Zhang, Z., Wang, X., Zhang, Z., Shen, G., Shen, S., and Zhu, W. Unsupervised graph neural architecture search with disentangled self-supervision. In *Advances in Neural Information Processing Systems*, 2023e.

Zhang, Z., Wang, X., Qin, Y., Chen, H., Zhang, Z., Chu, X., and Zhu, W. Disentangled continual graph neural architecture search with invariant modularization. In *International Conference on Machine Learning*, 2024a.

Zhang, Z., Wang, X., Zhang, Z., Li, H., Qin, Y., and Zhu, W. Llm4dyg: Can large language models solve spatial-temporal problems on dynamic graphs? In *Conference on Knowledge Discovery & Data Mining (ACM SIGKDD)*, 2024b.

A. Proof

A.1. Proof of Theorem 1

Proof. We prove that our learned graph representation is disentangled by showing it satisfies the equivariant property and decomposable property in Definition 1. The reasoning process is inspired by Lemma 1 in (Wang et al., 2021a), but we extend it into a more general scenario that the representation is disentangled *w.r.t.* $\prod_{k=1}^{K} \mathcal{G}^k$ instead of disentangled *w.r.t.* some $\mathcal{G}^k$ and the others in (Wang et al., 2021a), thanks to our tailored model design.

First, we prove that the graph representation learned from our method satisfies the equivariant property. Consider that the objective function Eq. (7) reaches the minimum for any k-th latent factor, $k \in \{1, \dots, K\}$. Assume that there are representations $\mathbf{Z}_n$ and $\mathbf{Z}_{n'}$ of two different graphs satisfying $\mathbf{Z}_n = \mathbf{Z}_{n'}$. The similarity between the two graphs can be further reduced by outputting different representations for the n-th graph and the n'-th graph for reaching a lower loss in Eq. (7). It contradicts that the objective function reaches the minimum. Therefore, the sample-equivariant property (Wang et al., 2021a) can be achieved, i.e., different graphs have different representations. We assume that the transition from the n-th graph to the n'-th graph in semantic space is from the group action and denote this action as $g \in \mathcal{G}$ that is transitive on representation space, i.e. $g \cdot \mathcal{Z}_n = \mathcal{Z}_{n'}$. Therefore, the sample-equivariant graph representations can be an approximation of group-equivariant graph representations (Wang et al., 2021a).

Then, we prove that the graph representation learned from our method satisfies the decomposability property. Consider that the objective function Eq. (7) reaches the minimum for any k-th latent factor, $k \in \{1, \dots, K\}$. Assume that $\exists \mathbf{Z}^k$ is affected by not only the action g^k but also one different action $g^{k'}$ and consider obtaining the environment partition by clustering the representation $\mathbf{Z}^{k'}$. The similarity between two graphs measured by the part $\mathbf{Z}^k$ can be further reduced by excluding the action $g^{k'}$ for reaching a lower loss in Eq. (7). Furthermore, since $\mathbf{Z}^k$ is also affected by $g^{k'}$ in addition to g^k, the invariant regularization term in Eq. (7) can also be further reduced by excluding the information of the k'-th latent factor. It contradicts that the objective function reaches the minimum. Finally, we conclude that each $\mathbf{Z}^k$ is affected only by g^k corresponding to the k-th latent factor and unaffected by the actions corresponding to the other latent factors.

Overall, the learned graph representation of our method can be disentangled. □

A.2. Proof of Theorem 2

Proof. The OOD generalization means that the model can make predictions only based on the invariant and sufficient disentangled graph latent factors. Therefore, given the disentangled graph representations $\mathbf{Z} = [\mathbf{Z}^1, \dots, \mathbf{Z}^K]$, the coefficient $\mathbf{w} = [w^1, \dots, w^K]$ of the linear predictor for predicting label Y can be regarded to identify the invariant and sufficient latent factors and the predictions can be OOD generalized with the corresponding disentangled graph representations. Following the reasoning line of Theorem 5.1 and Theorem 5.2 in (Xu et al., 2022), we assume that the ordinary least square method is adopted, the mean of each disentangled representation is zero, and the independence among the disentangled representations. We derive the solutions of the coefficient $\mathbf{w} = [w^1, \dots, w^K]$.

Specifically, we have

$$w^k = \frac{\mathrm{Cov}(\mathbf{Z}^k, Y)}{\mathrm{Var}(\mathbf{Z}^k)} = \frac{\mathbb{E}\left[\mathbf{Z}^k Y\right]}{\mathrm{Var}(\mathbf{Z}^k)} = \frac{\mathbb{E}\left[\mathbf{Z}^k \mathbb{E}\left[Y|\mathbf{Z}\right]\right]}{\mathrm{Var}(\mathbf{Z}^k)}, \tag{9}$$

where Cov and Var denote the covariance and variance respectively, and $\mathbb{E}$ denotes the expectation.

On the one hand, if k-th latent factor does not belong to the invariant and sufficient latent factors to predict the label Y, then $\mathbb{E}\left[Y|\mathbf{Z}\right]$ can be represented as the function $\Omega(\mathbf{Z}^{-k})$, where $\mathbf{Z}^{-k} = [\mathbf{Z}^1, \dots, \mathbf{Z}^{k-1}, \mathbf{Z}^{k+1}, \dots, \mathbf{Z}^K]$. Therefore,

$$w^k = \frac{\mathbb{E}\left[\mathbf{Z}^k \Omega(\mathbf{Z}^{-k})\right]}{\mathrm{Var}(\mathbf{Z}^k)} = 0. \tag{10}$$

The coefficient w^k is zero, meaning that the prediction does not rely on the representation corresponding to the k-th latent factor. The spurious correlation between this latent factor and the label can be removed, leading to OOD generalized predictions.

On the other hand, if k-th latent factor belongs to the invariant and sufficient latent factors to predict the label Y, then

$\mathbb{E}\left[Y|\mathbf{Z}\right]$ can be represented as the function $\Theta(\mathbf{Z}^k)$. Therefore,

$$w^k = \frac{\mathbb{E}\left[\mathbf{Z}^k \Theta(\mathbf{Z}^k)\right]}{\mathrm{Var}(\mathbf{Z}^k)} \neq 0. \tag{11}$$

The coefficient w^k is not zero, meaning that the prediction relies on the representation corresponding to the k-th latent factor. The invariant and sufficient correlation between this latent factor and the label can be identified accurately, leading to OOD generalized predictions.

□