Unraveling and Mitigating Endogenous Task-oriented Spurious Correlations in Ego-graphs via Automated Counterfactual Contrastive Learning

Tianqianjin Lin[a,b], Yangyang Kang[c], Zhuoren Jiang[a,*], Kaisong Song[b,f], Kun Kuang[d], Changlong Sun[b], Cui Huang[a], Xiaozhong Liu[e,*]

[a]*Department of Information Resources Management, Zhejiang University, P.R. China*
[b]*Institute for Intelligent Computing, Alibaba Group, P.R. China*
[c]*Polytechnic Institute, Zhejiang University, P.R. China*
[d]*College of Computer Science and Technology, Zhejiang University, P.R. China*
[e]*Computer Science Department, Worcester Polytechnic Institute, USA*
[f]*Northeastern University, P.R. China*

Abstract

Graph Neural Networks (GNNs) have been proven to easily overfit spurious subgraphs in the available data, which reduces their trustworthiness in high-stakes real-world applications. Current works often assumed that such spurious subgraphs are caused by a latent environment variable (e.g., selection bias) and addressed this issue by learning invariance across synthesized multiple environments. However, this work uncovers a prevalent yet overlooked mechanism in node-level tasks leading to spurious subgraphs without assumption on the environment variable. Moreover, the identified mechanism implies that the spurious subgraphs can differ in different tasks even within the same ego-graph. To mitigate this **E**ndogenous and **T**ask-oriented **S**purious **C**orrelations (ETSC), this work designs a novel and automated **C**ounterfactual **C**ontrastive **L**earning framework for **G**raphs in **n**ode-level tasks (CCL-Gn). Based on the analysis of the relationship between spurious subgraphs and causally correlated subgraphs to the task within this mechanism, we propose an original counterfactual optimization objective to separate them automatically and sufficiently. To further maintain a model-agnostic property, CCL-Gn enables GNN optimization with an auxiliary contrastive learning objective between the raw ego-graph and the counterfactual views. Extensive experiments on 13 datasets with 29 data splits demonstrate that CCL-Gn can consistently enhance the performance of a series of typical GNNs in both the in-distribution and out-of-distribution scenarios.

Keywords: node property prediction, counterfactual contrastive learning, graph contrastive learning, spurious correlations

1. Introduction

In a prediction task, spurious correlations represent the occurrence of unexpected correlations between a subset of input features and the target label, which are irrelevant to the causal mechanism of the task (Izmailov et al., 2022; Vigen, 2015).

Deep neural networks, due to their excessive fitting capacity or their tendency to learn shortcuts (Geirhos et al., 2020), can easily overestimate the significance of the spurious correlations (Sagawa et al., 2020; Fan et al., 2022a). Thereby, they can get compromised generalizability in unseen environments or trustworthiness in real-world applications (Yang and Chaudhuri, 2022). In the field of Graph Neural Networks (GNNs), many works have also revealed this problem (Li et al., 2022a; Miao et al., 2022; Knyazev et al., 2019; Lin et al., 2024). Hence, a thorough examination of spurious correlations within graph data becomes urgently important.

[*]Corresponding author

Email addresses: lintqj@zju.edu.cn (Tianqianjin Lin), yangyangkang@zju.edu.cn (Yangyang Kang), jiangzhuoren@zju.edu.cn (Zhuoren Jiang), kaisong.sks@alibaba-inc.com (Kaisong Song), kunkuang@zju.edu.cn (Kun Kuang), changlong.scl@taobao.com (Changlong Sun), huangcui@zju.edu.cn (Cui Huang), xliu14@wpi.edu (Xiaozhong Liu)

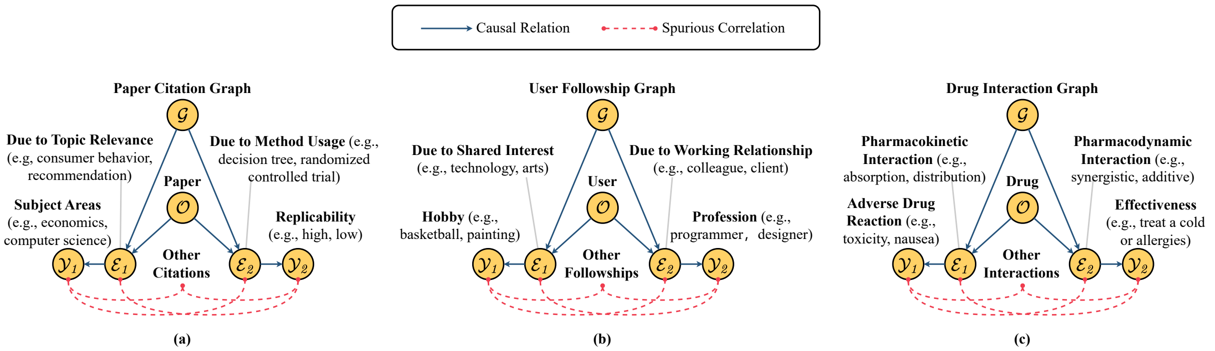

Causal Relation
Spurious Correlation
Paper Citation Graph
$\mathcal{G}$
Due to Topic Relevance (e.g, consumer behavior, recommendation)
Due to Method Usage (e.g., decision tree, randomized controlled trial)
Paper
$\mathcal{O}$
Subject Areas (e.g., economics, computer science)
Replicability (e.g., high, low)
$\mathcal{Y}_1$
$\mathcal{E}_1$
Other Citations
$\mathcal{E}_2$
$\mathcal{Y}_2$
(a)
User Followship Graph
Due to Shared Interest (e.g., technology, arts)
Due to Working Relationship (e.g., colleague, client)
User
Hobby (e.g., basketball, painting)
Profession (e.g., programmer, designer)
Other Followships
(b)
Drug Interaction Graph
Pharmacokinetic Interaction (e.g., absorption, distribution)
Pharmacodynamic Interaction (e.g., synergistic, additive)
Drug
Adverse Drug Reaction (e.g., toxicity, nausea)
Effectiveness (e.g., treat a cold or allergies)
Other Interactions
(c)

Figure 1: Toy examples of the Endogenous Task-oriented Spurious Correlations in ego-graph generation process. When a node ($\mathcal{O}$) is connected to a graph ($\mathcal{G}$), it often forms edges ($\mathcal{E}_1$ and $\mathcal{E}_2$) with different nodes within the graph based on various causes. When we define specific tasks ($\mathcal{Y}_1$ and $\mathcal{Y}_2$), these tasks typically have a causal relationship with only a subset of these edges ($\mathcal{E}_1 \rightarrow \mathcal{Y}_1$ and $\mathcal{E}_2 \rightarrow \mathcal{Y}_2$), while inevitably showing spurious correlations with other edges ($\mathcal{E}_1 \rightarrow \mathcal{Y}_2$ and $\mathcal{E}_2 \rightarrow \mathcal{Y}_1$) due to the node itself and the overall property of $\mathcal{G}$ acting as confounders for these edges. **(a)** Compared to papers in "computer science", papers in "economics" may have a higher proportion of citing randomized control trial. However, we should not judge the subject area of a paper by whether it employs the method of randomized control trial. For instance, papers in "computer science" involving A/B test in recommendation could also cite randomized control trial. Similarly, the replicability of research findings in a paper can vary across different topics, but the topic itself does not causally affect the replicability of the paper. The methodology mainly determine the replicability. **(b)** A user's hobbies causally influence the other users they follow based on shared interests, but there is no direct causal relationship with work-related connections. However, there is usually correlation between hobbies and work-related connections; for example, employees of an advertising design company may have a higher proportion of interest in painting compared to employees of an information technology company. **(c)** In pharmacy science, adverse drug reactions are more causally related to pharmacokinetic interactions, while the effectiveness of drugs for specific diseases is more causally related to pharmacodynamic interactions. However, they can exhibit spurious correlations; for example, drugs with strong pharmacodynamic interactions may show a higher risk of adverse reactions due to their strong combined effects, which is not because pharmacodynamic interactions directly cause adverse drug reactions, but because potent drugs themselves are more likely to cause noticeable adverse drug reactions.

Recently, invariant learning has been extended to node representation learning Wu et al. (2021); Li et al. (2023); Liu et al. (2023b) to address the issue of spurious correlations. This line of research usually assumes that an environment variable leads to variant or spurious correlations. Therefore, they attempt to construct multiple sub-environments within the training data to learn the invariance of node representations or task performances across different sub-environments Lin et al. (2022).

However, this work uncovers a prevalent yet overlooked inherent cause of spurious correlations in node-level graph tasks, which is irrelevant to the assumption on latent environment variable (e.g., selection bias). Moreover, this work suggests that the spuriously correlated structures can differ in different tasks even within the same ego-graph. In this paper, based on the above two properties of the mechanism, we refer to this type of spurious correlation as **E**ndogenous **T**ask-oriented **S**purious **C**orrelation (ETSC). This augments the understanding of spurious correlations and generalizability challenges in node-level tasks over graphs.

The detailed explanation is as follows. Firstly, node-level tasks are typically defined after graph formation, i.e., on naturally existing graphs Hu et al. (2020). In other words, the graphs are collected and saved without any particular scientific objectives and the tasks of interest are an afterthought. Consequently, the obtained ego-graphs include all observable edges caused by all the complex and diverse edge formation mechanism Jackson (2005); An et al. (2022), even though typically only a portion of the edges is causally related to the task of interest. Secondly, as the ego node acts as a confounder Pearl (2009) of different subsets of edges during the edge formation process, spurious correlations arise between the task and those non-causal edges. Obviously, when the tasks defined on the ego-graphs change, the causally related edges will change, and the non-causal edges will also change accordingly, resulting in the task-oriented property. We provide three toy examples across three domains in Figure 1. Unlike previous work that hypothesized the existence of a latent environmental variable leading to the contingency and static nature of spurious correlations, ETSC implies the inevitability and dynamic nature of spurious correlations in node-level graph tasks.

Due to the particularity of ETSC, previous work based on invariant learning also cannot optimally handle the ETSC. Since edge formation mechanism remains relatively consistent across the entire graph, spuriously correlated edges can be universal and maintain a certain level of predictive invariance in divided sub-environments, making it difficult

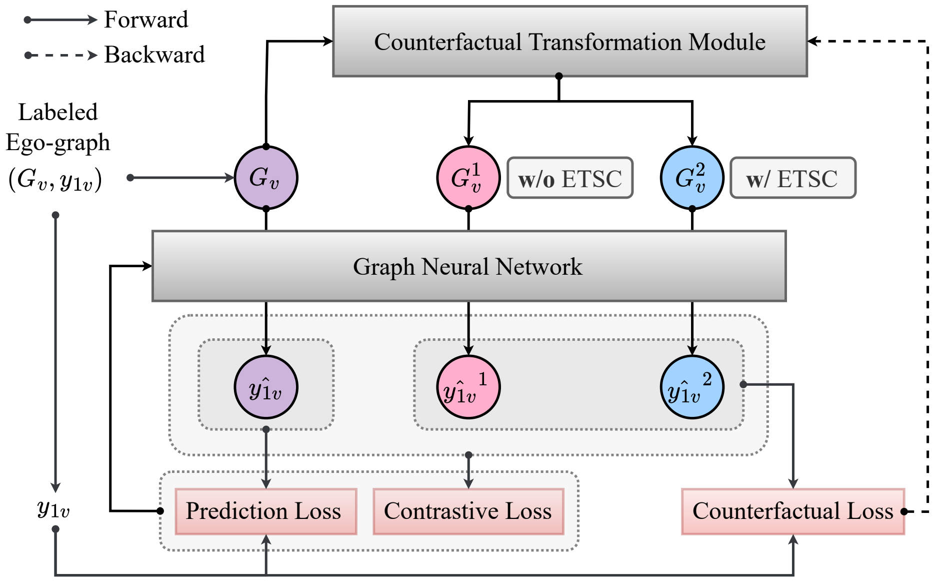


Figure 2: Overall Architecture of the Proposed CCL-Gn.

to be separated sufficiently from the invariant perspective. Furthermore, due to the diversity of the edge formation mechanism, the ETSC can exhibit strong heterogeneity, for example, time- or degree-driven attachment Topirceanu et al. (2018). When the basis for dividing the environment, whether it's prior knowledge or optimization obejctives, is relatively singular, these methods may not yield optimal solutions or may be limited in their generalizability across different tasks and data splits Tan et al. (2023); Yang et al. (2023b); Chen et al. (2023a).

Moreover, it's observed that the ETSC can even exist in the typical in-distribution setting. While it is potentially useful for predictions in both training and testing sets, models built on them may not align with expectations in real-world applications Ying et al. (2019); Fan et al. (2022a); Lin et al. (2024). In fact, real-world graphs usually have limited labeled nodes Dai et al. (2022), as a result, it can even affect performance on the in-distribution test set due to the issue of overfitting Esser et al. (2021).

To deal with the ETSC, our focus has shifted away from environment and distributional nuances. Instead, we redirect our attention towards the intrinsic characteristics of the ego-graph itself and introduce a novel model-agnostic framework named Counterfactual Contrastive Learning for Graphs in node-level tasks (CCL-Gn), as shown in Figure 2. Based on the analysis of the relationship between the causally relevant structures and the ETSC within the ego-graph in node-level tasks, CCL-Gn learns to decompose the ego-graph into the ego-subgraph causally correlated to the task (i.e., ego-subgraph without ETSC) and the ego-subgraph spuriously correlated to the task (i.e., ego-subgraph with ETSC) via an original and theoretically-guaranteed counterfactual optimization objective. Subsequently, CCL-Gn forces GNNs to focus on causal ego-subgraph and enables GNNs to mitigate the ETSC by leveraging contrastive learning techniques He et al. (2020); Gao et al. (2021b); You et al. (2020). The raw ego-graph is pulled closer to the causal ego-subgraph and pushed apart from the noncausal ego-subgraph in the representation or label probability space.

The core contribution of this paper can be threefold:

1. We introduce a novel perspective on the origins of spurious correlations in node-level tasks, i.e., ETSC, thereby providing a valuable complement to the research on generalizatable node representation learning.
2. We design a novel model-agnostic framework, i.e., CCL-Gn, to intentionally address ETSC. It is capable of automatically identifying the edges of the ETSC from the input ego-graph under specific tasks.
3. We conduct extensive experiments on 13 real-world datasets with 29 data splits to validate the superiority of

CCL-Gn over state-of-the-art Graph Contrastive Learning (GCL) and graph out-of-distribution (OOD) methods in terms of improving both the performance and generalizability of typical GNNs.

2. Related Works

This work is motivated by the spurious correlations evident within the ego-graph, prompting the development of an innovative counterfactual graph contrastive learning paradigm. Hence, we discuss three categories of related works.

Out-of-distribution (OOD) Generalization in Graphs. The rapid advancement of GNNs in practical scenarios has brought increased interest in the generalizability of GNNs. Despite the impressive progress made in graph-level tasks Wu et al. (2022); Fan et al. (2022b); Li et al. (2022b); Chen et al. (2022); Zhuang et al. (2023), the node-level tasks have not been as extensively studied. Current methods designed for node-level tasks mostly follow the well-established principles of invariant learning Arjovsky et al. (2019). The key focus of them is on how to create various distinct environments. For instance, EERM Wu et al. (2021) and FLOOD Liu et al. (2023b) utilize learnable graph editors to generate graphs as virtual environments and maximize the variance of risks across the environments. INL Li et al. (2023) constructs different environments by clustering the graph through modularity maximization. However, the efficacy of environment construction remains uncertain without additional information Lin et al. (2022); Yang et al. (2023b). Moreover, the optimal node representations may even need to adapt to different environments instead of keeping invariant Liu et al. (2023a). In addition to invariant learning, some efforts originate from the perspectives of data augmentation Wang et al. (2021) or domain adaptation Ganin et al. (2016); Jin et al. (2022); Liu et al. (2023a), possessing a certain ability to address OOD issues as well. However, due to the goal shift, the effects obtained by these efforts are often limited Gui et al. (2022); Wu et al. (2023). As a comparison, our method doesn't require partitioning environments or the assumption of distribution shift.

Graph Contrative Learning (GCL). To learn generalizable, transferable and robust representations for graph data (You et al., 2021), GCL (You et al., 2020; Hassani and Ahmadi, 2020; Zhu et al., 2021b; Li et al., 2022c; Xia et al., 2023) has been extensively studied recently. They optimize GNNs via contrastive objectives to discriminate positive graph pairs from negative graph pairs (Xie et al., 2022; Liu et al., 2022). In these works, two types of GCL are more relevant to this research. One is supervised GCL Khosla et al. (2020); Yin et al. (2022); Peng et al. (2024), which generates contrastive views under task supervision to keep task-relevant information intact in the positive view. Because the ETSC is predictive, the trained GNNs under this approach are prone to fit the ETSC as well, similar to vanilla GNN training. The other is rational/saliency-based GCL Li et al. (2022c); Wei et al. (2023); Yang et al. (2023a); Chen et al. (2023b). This approach aims to identify a graph's most semantically discriminative structures through contrastive learning. However, these methods are all designed for graph-level tasks. Moreover, for self-supervision, the ETSC itself is also crucial for the ego node to keep semantics distinguishable. They only exhibit spurious correlations when specific tasks are considered. Therefore, the representations of nodes under this objective will still contain subgraphs of spurious correlations for a given task. In summary, the current GCL cannot effectively identify and separate the ETSC.

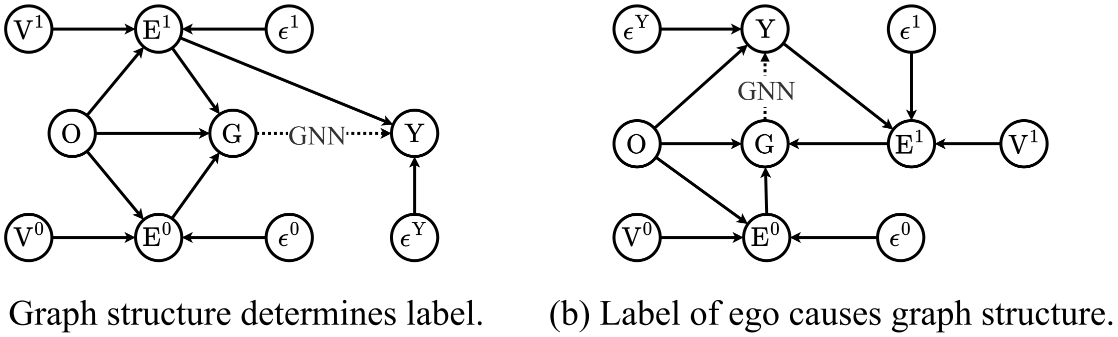
Counterfactual Contrastive Learning (CCL). In the text and image fields, the CCL has been proposed to alleviate the negative impact of spurious correlations. This approach first identifies causal features (to a specific learning problem) in raw samples. Then it creates pairs of positive/negative counterfactual views by retaining/removing the identified parts. Leveraging contrastive learning techniques, CCL further learns a triplet relationship: the raw samples are pulled closer to the positive counterfactual views and pushed apart from the negative counterfactual views (Liang et al., 2020; Zhang et al., 2021b). Consequently, CCL forces models to focus on causal features in the input samples and enables models to achieve promised generalizability in unseen data. Various real-world applications, e.g., visual question answering (Zhang et al., 2020; Liang et al., 2020; Chen et al., 2020; Shu et al., 2024), text classification (Choi et al., 2022; Fan et al., 2024) and sequence recommendation (Zhang et al., 2021b), have validated the efficiency and superiority of this approach. However, prior CCL works in the text/image domain primarily rely on rules or manual annotating to generate counterfactual views. Unlike texts/images, real-world graphs are more complex and abstract (Zhou et al., 2020; Gao et al., 2021a). It's hard for humans to interpret intuitively/visually and identify the causal parts. Therefore, CCL in the graph domain is still underexplored.

3. Method

3.1. Preliminaries

In this section, we will first illustrate the Endogenous Task-oriented Spurious Correlations (ETSC) based on causal diagrams Pearl (2009) of the 1-hop ego-graph generation process. Based on the analysis, we will then give formal definitions involved in CCL-Gn and formulate the key research questions.

Endogenous Task-oriented Spurious Correlations. To begin with, we divide node-level tasks into two categories, as shown in Figure 3.



(a) Graph structure determines label.

(b) Label of ego causes graph structure.

Figure 3: ϵ represents random noise. Solid arrows represent causal relationships. The dotted line with text GNN represents that the whole observed graph G is generally used to predict the task label Y of the ego O. **(a)** The ego O can cause two kinds of edges: E^0 (with nodes V^0) and E^1 (with nodes V^1). Only E^1 determines the label Y. **(b)** The ego O itself determines the label Y. Y causes E^1 (with nodes V^1), while O cause E^0 (with nodes V^0).

In both cases, the relationship between E^1 and Y can be relatively stable since E^1 is the direct cause or effect of Y. However, the joint distribution of E^0 with Y would be different if the ego set changes since E^0 and Y are confounded by ego O (Schölkopf et al., 2012). GNNs trained without any intervention inevitably fit the spurious correlations between G^- and Y since they take the whole observed graph G as input. Analysis of k-hop ($k > 1$) ego-graph generation process and interpretative examples are provided in our Appendix A and B. Here, we give the formal definition of Endogenous Task-specific Spurious Correlations.

Definition 3.1 (Endogenous Task-specific Spurious Correlations, ETSC)**.** Given a graph $G = (V, E)$, where V is the set of nodes and E is the set of edges, suppose that each edge $e_{uv} \in E$ between nodes $u \in V$ and $v \in V$ exists due to one or more of T mechanisms, represented by the set $F = \{f_1, f_2, \dots, f_T\}$. Each mechanism $f_i : (u, v) \to \{0, 1\}$ is a binary function indicating whether an edge between nodes u and v is generated due to mechanism f_i. An edge e_{uv} is observed if and only if $\sum_{i=1}^{T} f_i(u, v) \geq 1$, meaning that at least one mechanism f_i leads to the creation of e_{uv}. Let the set of all non-empty proper subsets of F be denoted as $\mathcal{P}^*(F) = \{f' \mid f' \subset F, f' \neq \emptyset, f' \neq F\}$. Now, consider a node-level task where the goal is to predict a property Y of a node $o \in V$. Suppose that Y is causally influenced only by edges generated by mechanisms in a subset $F_Y \in \mathcal{P}^*(F)$. This causal influence may follow one of two possible directions: either (i) edges generated by mechanisms in F_Y causally influence Y, denoted by $\{f_i(o, u) \mid u \in V, f_i \in F_Y\} \to y_o$, or (ii) Y causally influences the existence of these edges, denoted by $y_o \to \{f_i(o, u) \mid u \in V, f_i \in F_Y\}$, where $\to$ indicates the direction of causation. The mechanisms in the complementary set $F_{ETSC} = F \setminus F_Y = \{f_i \in F \mid f_i \notin F_Y\}$ are considered to be **Endogenous Task-specific Spurious Correlations (ETSC)** in node-level graph tasks.

Motivated by the CCL research, we further conclude the E_1 and E_2 here as the definitions of positive and negative counterfactual samples as follows.

Definition 3.2 (Optimal Task-oriented Positive Counterfactual View G^+)**.** An ego-subgraph $G^{'}$ is the optimal task-oriented positive counterfactual view G^+ to an ego-graph G if and only if $E^{'}$ contains and only contains all the edges that are causally correlated to the task label (i.e., E_1).

Definition 3.3 (Optimal Task-oriented Negative Counterfactual View G^{-})**.** An ego-subgraph $G^{'}$ is the optimal task-oriented negative counterfactual view G^{-} to an ego-graph G if and only if $G^{'}$ is the complement of the optimal task-oriented positive counterfactual view G^{+} (i.e., E_2).

To force the GNNs to focus on structures that are causally relevant to the task without changes in model structure and inference flow, we can train GNNs with an auxiliary contrasting learning objective with the optimal task-oriented counterfactual views, i.e., G^{+} and G^{-}. Our goal is to make the learned representation of G to be as independent as possible from G^{-}, so that the model can achieve better generalizability. Let $\phi(G) : \{G \mid v_i \in \mathcal{V}\} \rightarrow \mathbb{R}^c$ be any GNN that maps an ego-graph G to a probability distribution $\boldsymbol{p}$ of the ego over the label space $\mathbb{R}^c$. The contrastive learning objective is defined as:

Definition 3.4 (Counterfactual Graph Contrastive Learning)**.** Given a G with its G^{+} and G^{-}, learning a function ϕ that maximizes the consistency between pair $(\phi(G), \phi(G^{+}))$ compared with pair $(\phi(G), \phi(G^{-}))$.

Research Questions. Since the counterfactual view is the prerequisite, the key challenge is ***Q1:*** *How to automatically generate G^{+} and G^{-}?* After we obtain the counterfactual views, to assess their quality, we should further investigate ***Q2:*** *Can we really enhance the GNN ϕ on given tasks with the counterfactual graph contrastive learning?*

3.2. Counterfactual Transformation Module

To answer Q1, in this section, we introduce the Counterfactual Transformation Module (CTM) φ. It estimates the probability that an edge is part of the optimal task-oriented counterfactual positive view G^{+} for a G. We will first describe the forward process of the estimation and then explain the proposed counterfactual loss.

Representation of an Edge. We represent an edge (u, ς) in G_v by the concatenation of the raw node features of the ego node v, the source node u, and the destination node ς:

$$\boldsymbol{e}_{u\varsigma|G_v} = \boldsymbol{x}_v \parallel \boldsymbol{x}_u \parallel \boldsymbol{x}_\varsigma. \tag{1}$$

Considering the specificity of the ego-graph setting, we additionally add the structural information in the representation of the edge. Technically, for a node u in an ego-graph G_v, we can assign an ego identifier defined as $\mathbf{1}_{u=v}$ and then construct auxiliary structural information leveraging Distance Encoding (DE) Li et al. (2020) defined as landing probabilities of random walks of different lengths from ego v to the node u, which is denoted by $\boldsymbol{r}_{u|G_v}$. Therefore, a new node feature vector for u can be obtained as

$$\tilde{\boldsymbol{x}}_{u|G_v} = \boldsymbol{x}_u \parallel \boldsymbol{r}_{u|G_v} \parallel \mathbf{1}_{u=v}, \tag{2}$$

where $\boldsymbol{r}_{u|G_v} = [(\boldsymbol{A}\boldsymbol{D}^{-1})_{u,v}, (\boldsymbol{A}\boldsymbol{D}^{-1})^2_{u,v}, \ldots, (\boldsymbol{A}\boldsymbol{D}^{-1})^l_{u,v}]$. $\boldsymbol{A}$ is the adjacency matrix for G_v such that $\boldsymbol{A}_{u,\varsigma} = 1$ iff $(u, \varsigma) \in E_v$ and $\boldsymbol{D}$ is the degree (diagonal) matrix for G_v where $\boldsymbol{D}_{u,u}$ is the degree of node u. l indicates the length of random walks, which is set to the same as the hop k of the ego-graph. Then an edge (u, ς) in G_v can be represented as

$$\boldsymbol{e}_{u\varsigma|G_v} = \tilde{\boldsymbol{x}}_{v|G_v} \parallel \tilde{\boldsymbol{x}}_{u|G_v} \parallel \tilde{\boldsymbol{x}}_{\varsigma|G_v}. \tag{3}$$

Probability That an Edge Belongs to G_v^{+}. For simplicity, we assume the binary variable $\mathbf{1}_{(u,\varsigma)\in E_v^{+}}$ follows a Bernoulli distribution $\mathrm{Ber}(\theta_{u\varsigma|G_v})$ and $\theta_{u\varsigma|G_v}$ is estimated by multiple multilayer perceptrons (MLPs) as follows:

$$P\left((u, \varsigma) \in E_v^{+}\right) = \theta_{u\varsigma|G_v} = \mathrm{S}\left(\frac{1}{m}\sum_{i=1}^{m}\mathrm{MLP}_i\left(\boldsymbol{e}_{u\varsigma|G_v}; \Lambda_i\right)\right), \tag{4}$$

where S is the sigmoid and m is set to 4 for simplicity.

Assuming all the edges are independent of each other, the probability of a sampled sub-egograph $G_v^{'}$ being G_v^{+}, i.e., $P(G_v^{+} = G_v^{'})$ can be obtained by

$$\begin{aligned} P(G_v^{+} = & G_v^{'}) = \Pi_{(u,\varsigma)\in E_v^{'}} P\left((u, \varsigma) \in E_v^{+}\right) \\ & \times \Pi_{(u,\varsigma)\in E_v \setminus E_v^{'}} \left(1 - P\left((u, \varsigma) \in E_v^{+}\right)\right). \end{aligned} \tag{5}$$

Note that once G_v^+ has been obtained, we can further construct $G_v^- = \bar{G}_v^+$. Due to the discrete nature, we adopt the reparameterization trick Jang et al. (2017); Maddison et al. (2017) to enable updating parameters in the MLPs with general gradient-based optimizer (please refer to Appendix E).

Counterfactual Loss. To satisfy the signature of G_v^+ in Definition 3.2, the minimum negative log-likelihood loss (NLL) $\mathcal{L}_{pred}$ on $\phi^+(G_v^+; \Theta^+)$ can be minimized to encourage G_v^+ sufficient to predict the correct label y_v as follows:

$$\begin{aligned} \mathcal{L}_{suff}(\varphi \mid \hat{\phi}^+) =& \mathbb{E}_{G_v^+ \sim P(G_v^+ = G_v')} \left[\mathcal{L}_{pred}(\hat{\phi}^+) \right], \\ & \text{where } \hat{\phi}^+ = \arg\min_{\phi^+} \mathcal{L}_{pred}(\phi^+). \end{aligned} \tag{6}$$

However, without constraints on this loss, $\hat{\phi}^+$ tends to include edges as much as possible to achieve a lower training loss. Hence, to enforce φ to find the most significant edges, G_v^+ should be as sparse as possible, which aligns with the findings of previous studies in causal discovery and counterfactual explanation (Zheng et al., 2018; Wachter et al., 2017). A L1 regularization $\mathcal{L}_{size}$ is applied on the probability of being selected into E_v^+ for all edges:

$$\mathcal{L}_{size}(\varphi) = \mathbb{E}_{\forall (u,\varsigma) \in E_v^+} \left[P\left((u, \varsigma) \in E_v^+ \right) \right]. \tag{7}$$

Because G_v^- also has predictive power due to the spurious correlation, it will be potentially included in the G_v^+ if only $\mathcal{L}_{suff}$ is considered, even under the constraint of $\mathcal{L}_{size}$. To establish effective supervision signals, we investigated the relationship between the predictive power of G_v^+ and G_v^- and discovered a rank relationship between them:

Proposition 3.5. *Given a set* $\{(G_v,\ y)\}$, $\hat{\phi}^-$ *estimated with* $\{(G_v^-,\ y)\}$ *suffers a greater error rate compared to* $\hat{\phi}^+$ *estimated with* $\{(G_v^+,\ y)\}$.

The proof and discussion are left to the Appendix C. Therefore, under Proposition 3.5, a new optimization objective can be established by a rank loss $\mathcal{L}_{rank}(\varphi \mid \hat{\phi}^+, \hat{\phi}^-)$ that enforces $\hat{\phi}^-(G_v^-; \hat{\Theta}^-)$ to have greater empirical loss (than $\hat{\phi}^+$) with sufficient predictive power, i.e.,

$$\mathcal{L}_{rank}(\varphi \mid \hat{\phi}^+, \hat{\phi}^-) = \mathbb{E}_{G_v^+ \sim P(G_v^+ = G_v')} \left[\frac{\mathcal{L}_{pred}(\hat{\phi}^+)}{\mathcal{L}_{pred}(\hat{\phi}^-)} \right], \tag{8}$$

where $\hat{\phi}^- = \arg\min_{\phi^-} \mathcal{L}_{pred}(\phi^-)$. The choice of the function form is discussed in Appendix D. Unlike prior works on learnable augmentation or graph explanation, we do not directly maximize the empirical risk of ϕ estimated with $\{(G^-,\ y)\}$, highlighting that the G^- is also predictive to the label as analyzed.

The magnitude of the predictive power of the spurious structures G_v^- varies across graph and task settings. Therefore, we introduce a hyperparameter α to control the value of $\mathcal{L}_{rank}$. Moreover, the prior proportions of the causally-correlated structures (i.e., $\frac{E_v^+}{E_v^+ + E_v^-}$) can be also different in different graphs and tasks. Therefore, we additonaly add another hyperparameter β to control $\mathcal{L}_{size}$. Finally, the counterfactual loss for the counterfactual transformation module can be formulated as:

$$\mathcal{L}_{cf}(\varphi) = \mathcal{L}_{suff} + \alpha \cdot \mathcal{L}_{rank} + \beta \cdot \mathcal{L}_{size}. \tag{9}$$

We have conducted sensitivity experiments of coefficient α and β, as well as the ablation studies in the Section 4.4.

3.3. Counterfactual Graph Contrastive Learning

Contrasting the raw ego-graphs to both the positive and negative counterfactual views can inject task knowledge from the counterfactual transformation module φ into the GNNs ϕ. This will enhance the model performance and generalizability. An InfoNCE (van den Oord et al., 2018) style loss is formulated as

$$\mathcal{L}_{cl}(\phi) = \mathbb{E}_{\{(G_v, G_v^+, G_v^-) \mid \forall v \in \mathcal{V}\}} \mathrm{CL}(G_v, G_v^+, G_v^-; \tau), \tag{10}$$

where $\mathrm{CL}(G_v, G_v^+, G_v^-; \tau)$ is given as

$$-\log\left(\frac{\exp\left(\mathrm{sim}(G_v, G_v^+)/\tau\right)}{\exp(\mathrm{sim}(G_v, G_v^+)/\tau) + \exp\left(\mathrm{sim}(G_v, G_v^-)/\tau\right)}\right), \tag{11}$$

where $\mathrm{sim}(G_v, G_v^{'}) = 1 - \mathrm{JSD}(\phi(G_v), \phi(G_v^{'}))$ and JSD is the Jensen Shannon divergence. τ is the temperature parameter and plays a role in controlling the strength of penalties on the task-oriented counterfactual negative views. The larger the τ is, the smaller the influence of similarity between original ego-graphs and the task-oriented counterfactual negative views is.

3.4. Joint Learning of GNN and CTM

For simplicity and efficiency, we empirically share ϕ with ϕ^+ and ϕ^-, and jointly learn ϕ and φ with the counterfactual contrastive learning. Therefore, a final joint loss can be formulated as:

$$\mathcal{L}_{joint}(\phi, \varphi) = \mathcal{L}_{pred}(\phi) + \mathcal{L}_{cf}(\varphi \mid \phi) + \mathcal{L}_{cl}(\phi). \tag{12}$$

We have conducted an ablation study on the $\mathcal{L}_{cl}(\phi)$ loss in the Section 4.4.

3.5. Computational Complexity

In the counterfactual transformation module, the complexity for the edge representation $\boldsymbol{e}_{u\varsigma|G_v}$ is $O(l \cdot |V_v| \cdot |E_v|)$ due to the computation of $\tilde{\boldsymbol{x}}_{u|G_v}$ in Equation 2 (Yuster and Zwick, 2005) and the complexity for the probability estimation is $O((1 + \frac{h-1}{m}) \cdot |E_v| \cdot d^2)$. $|V_v|$ and $|E_v|$ are node size and edge size in the ego-graph of a node. l is the length of random walks, which is set to the same as the hop k of the ego-graph. d is the hidden dimension of node features and m and h are the number of MLPs and the number of layers of MLPs in equation 4 respectively. Here, as a comparison, the complexity for GNN-based flow used in learnable views generation (Yin et al., 2022; Li et al., 2022c) is $O(k \cdot |E_v| \cdot C_e + k \cdot |V_v| \cdot C_v + h \cdot |E_v| \cdot d^2)$ where C_e and C_v are the complexity of message passing and combination respectively. In practice, the latter is likely to be greater than our approach due to large C_e and C_v.

In the counterfactual contrastive learning module, CCL-Gn does not need to contrast in-batch negative samples since it only contrasts raw graphs with their negative counterfactual views. This helps CCL-Gn have time and space complexities scaling linearly to the batch size while the time and space complexities of most GCL methods are proportional to the square of the batch size. Furthermore, We conducted real runtime experiments in Appendix K. The results show that CCL-Gn achieves a very fast speed compared to other baselines. Especially when the graph becomes larger and more complex, e.g. on the Flickr dataset, CCL-Gn can achieve the fastest speed.

4. Experiments

To answer Q2, we first constructed a synthetic graph based on the motivation, i.e., ETSC, to illustrate the functionality of the CCL-Gn. Then, we investigate graph OOD benchmark datasets to validate the ability of CCL-Gn to enhance GNN generalizability, and compared CCL-Gn with various Graph OOD and GCL methods on various real-world datasets. Due to the page limitation, we leave the details of datasets and baselines in the Appendix F and G.

4.1. An Explanatory Experiment

We assume that nodes on the graph have two types of characteristics: color and shape. The edge formation mechanism in this graph is that nodes tend to connect with others that have the same color or shape. Node features are sampled from a multivariate normal distribution parameterized by the two characteristics. Two tasks are defined as predicting a node's color or shape based on the node's 1-hop ego-graph.

We found that CCL-Gn can generate positive and negative counterfactual contrastive views that are very close to the ideal state, as shown in Figure 4. The optimized GAT with CCL-Gn reduces the attention weights on neighbors with spurious correlations. Specific synthesis processes and experimental results are in the Appendix F.1

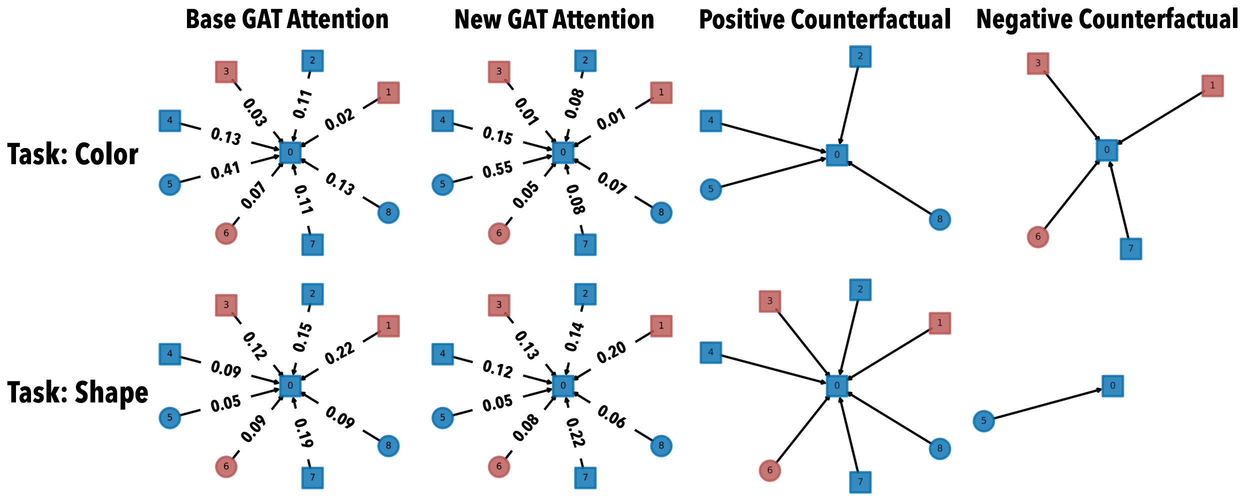


Figure 4: A real case from the explanatory experiment. CCL-GN generated counterfactual views according to the learning task.

4.2. Comparison with Graph OOD Algorithms

Based on the intention of addressing spurious correlation, we first compare CCL-Gn with a series of OOD algorithms to assess its capabilities in causal representation learning.

Dataset. Three datasets from the GOOD benchmark Gui et al. (2022), i.e., CBAS, WebKB and Cora, are included. Moreover, following (Wu et al., 2021; Li et al., 2023), we conduct a test on OGBN-Arxiv, where the distribution shifts are caused by selecting papers published before 2011 as training set, within 2011–2014 as validation set, and within 2014–2016/2016–2018/2018–2020 as three testing sets.

Baselines. We compare CCL-Gn with IRM Arjovsky et al. (2019), VREx Krueger et al. (2021), GroupDRO Sagawa et al. (2019), DANN Ganin et al. (2016), DeepCoral Sun and Saenko (2016), Mixup Wang et al. (2021), SRGNN Zhu et al. (2021a), EERM Wu et al. (2021), FLOOD Liu et al. (2023b), GIL Li et al. (2022b) and INL Li et al. (2023).

As shown in Table 1 and Table 2, CCL-Gn can outperform current state-of-the-art graph OOD methods or achieve comparable performances across all the settings. This validates its promising ability to enhance the GNN's generalizability and also demonstrates its own generalizability across tasks and distribution shifts. Remarkably, CCL-Gn outperforms the best baseline, i.e., EERM, on WebKB Covariate by 18%.

4.3. Comparison with GCL Algorithms

GCL often allows the representations learned by the model to be more robust and generalizable, thereby being validated as an auxiliary learning objective to enhance model performance in supervised scenarios Xie et al. (2022). Because CCL-Gn is also a form of contrastive learning, comparing it with the GCL works can verify the quality of the counterfactual contrastive views generated by CCL-Gn.

Dataset. Ten open benchmark datasets across three domains are investigated: (a) seven citation networks (Bojchevski and Günnemann, 2018) including CoraFull, CoraML, CiteSeer, DBLP, PubMed, Ogbn-Arxiv and Ogbn-MAG. (b) two product networks (Shchur et al., 2018) including Computers and Photo. (c) an image network (Zeng et al., 2020) Flickr.

Baselines. We compare CCL-Gn with three groups of GCL methods: (a) GCL with uniform/adaptive data augmentation, including DGI (Velickovic et al., 2019), MVGRL(Hassani and Ahmadi, 2020), Grace(Zhu et al., 2020), GCA(Zhu et al., 2021b), BGRL(Thakoor et al., 2021) and CCA-SSG (Zhang et al., 2021a); (b) GCL based on rational/saliency, including RGCL(Li et al., 2022c), CGC (Yang et al., 2023a) and GCIL (Mo et al., 2024); (c) supervised GCL, AutoGCL(Yin et al., 2022).

In-distribution Setting. As shown in Table 3, CCL-Gn can significantly enhance the performance of base GNNs. Specifically, CCL-Gn improves the accuracy of the base GAT from 1.13% (on Computers) to 9.19% (on Flickr).

Table 1: Comparison with Graph OOD Algorithms: Node classification accuracy on the three datasets from the GOOD benchmark. Results of FLOOD are reported in Liu et al. (2023b) and results of other baselines are from Gui et al. (2022). ERM refers to the basic empirical risk minimization. Base model is a 3-layer GCN, with the same configuration as in GOOD. **Bold** and underline represents the best and the second best performances, respectively.

Dataset	GOOD-CBAS		GOOD-WebKB		GOOD-Cora				avg.
Domain	Color		University		Word		Degree		
Shift	Covariate	Concept	Covariate	Concept	Covariate	Concept	Covariate	Concept	
ERM	76.00±3.00	82.36±0.97	14.29±3.24	27.83±0.76	64.86±0.38	64.60±0.17	56.30±0.49	60.54±0.44	55.85
IRM	76.00±3.39	83.21±0.54	13.49±0.75	27.52±0.43	64.77±0.36	64.60±0.16	56.28±0.63	61.23±0.32	55.89
VREx	77.14±1.43	82.86±1.26	14.29±3.24	27.83±0.38	64.80±0.28	64.57±0.18	56.30±0.50	60.58±0.42	56.05
GroupDRO	76.14±1.78	82.00±1.46	17.20±0.76	28.14±1.12	64.72±0.34	64.62±0.17	56.29±0.43	60.65±0.31	56.22
DANN	77.57±2.86	82.50±0.72	15.08±0.37	26.91±0.63	64.77±0.42	64.51±0.19	56.10±0.59	60.78±0.38	56.03
DeepCoral	75.86±3.06	82.64±1.40	13.76±1.30	28.75±1.13	64.72±0.36	64.58±0.18	56.35±0.38	60.58±0.40	55.91
Mixup	70.57±7.41	64.57±1.81	17.46±1.94	31.19±0.43	65.23±0.56	64.44±0.10	**58.20±0.67**	**63.65±0.39**	54.41
SRGNN	74.29±4.10	81.43±0.34	13.23±2.93	27.52±0.43	64.66±0.21	64.62±0.07	54.78±0.10	61.08±0.09	55.20
EERM	53.86±13.75	64.29±0.00	24.61±4.86	27.83±4.12	61.98±0.10	63.09±0.36	56.88±0.32	58.38±0.04	51.37
FLOOD	**83.53**	84.25	18.95	31.95	**66.23**	65.23	56.64	63.64	58.80
CCL-Gn	77.14±2.67	**85.14±1.06**	**29.05±6.36**	31.19±1.72	66.07±0.14	**65.44±0.32**	57.27±0.50	63.46±0.47	**59.35**

Table 2: Comparison with Graph OOD Algorithms: Node classification accuracy on the three test sets in OGBN-Arxiv. Results of baselines refer to Li et al. (2023). Base model is a 2-layer GAT, with the same configuration as in INL.

	OGBN-Arxiv		
Test Set	2014-2016	2016-2018	2018-2020
ERM	45.94±1.03	43.52±0.95	40.42±0.98
IRM	46.73±0.91	44.32±0.91	42.04±0.99
GroupDRO	45.95±0.89	43.52±1.25	40.43±1.32
VREx	45.93±0.87	45.69±0.81	41.01±1.03
EERM	45.99±1.22	45.32±0.84	42.01±1.36
GIL	47.70±0.93	45.65±1.41	41.87±1.89
INL	50.37±1.01	**49.12±1.23**	45.35±1.32
CCL-Gn	**50.53±0.23**	48.84±0.34	**45.80±0.42**

Considering all settings, the average gain of CCL-Gn for the base GAT is around 3.14%. Morever, CCL-Gn can consistently outperform the SOTA methods from 0.06% (on Computers) to 1.8% (on DBLP) except for setting of Flickr, where CCL-Gn got the second best performance. Note that on datasets like Photo, though the SOTA methods have already achieved a high performance (>93%), CCL-Gn can still push the boundary forward (>94%). All these results can prove the ability of CCL-Gn to enhance model performance by learning high-quality counterfactual views. Meanwhile, GCA, an approach that adopts adaptive data augmentation based on network theories and thus enables models to learn important structures, outperforms other baselines in most settings. This can also imply the effectiveness of CCL-Gn to auto-select the causally relevant structures from the network formation perspective.

Artificial Out-of-distribution. To provide empirical evidence for the causal analysis of Figure 3, we further conduct experiments based on the confounder, i.e., the ego. Following He et al. (2022), for each dataset, we run a Node2Vec (Grover and Leskovec, 2016) to get node embeddings and cluster the nodes into two clusters by K-Means (Lloyd, 1982). The cluster with a larger sample size is randomly divided into training and validation sets; the other is regarded as a testing set.

While the detailed results are left in our Appendix I.3, we show an interesting and exciting result in Figure 5: *The*

Table 3: Comparison with GCL Algorithms: Node classification accuracy on independently and identically distributed test sets. All the GCL methods (include CCL-Gn) function as an auxiliary learning objective as suggested in Xie et al. (2022). A 2-layer GAT was selected as base GNN. OOM indicates out-of-memory. The results of GraphSAGE or GIN as the base model are in the Appendix I.2.

Method	CiteSeer	Computers	Cora_Full	Cora_ML	DBLP	Flickr	Photo	PubMed
ERM	90.19±0.51	89.60±0.38	58.11±1.19	83.56±1.39	81.20±0.58	47.31±4.31	92.88±0.71	85.25±0.70
DGI	90.79±0.63	88.75±1.14	56.99±0.55	82.34±1.24	82.32±0.69	48.05±0.60	91.79±0.45	85.45±0.37
MVGRL	88.64±0.39	OOM	OOM	81.96±1.11	81.25±0.77	OOM	OOM	86.10±0.81
Grace	89.77±0.33	90.15±0.41	59.28±0.74	84.22±2.21	82.52±0.49	51.56±0.29	93.32±0.45	86.34±0.62
GCA	90.22±0.71	90.56±0.45	59.62±0.52	84.49±0.73	82.60±0.54	**51.91±0.69**	93.49±0.68	86.75±0.27
BGRL	90.09±1.17	89.46±0.88	58.13±0.78	84.44±1.29	81.52±0.83	50.55±1.12	92.50±0.70	85.99±0.40
CCA-SSG	90.95±0.67	89.72±1.01	60.15±0.89	84.88±0.47	83.86±0.45	49.89±1.42	92.43±0.94	86.46±0.48
RGCL	88.94±1.56	61.64±23.22	45.60±0.92	79.64±1.58	80.06±1.55	42.38±0.24	86.56±3.32	80.83±1.93
CGC	90.40±0.34	OOM	OOM	84.53±1.78	OOM	OOM	OOM	85.90±0.42
GCIL	90.84±0.54	90.04±0.80	60.42±0.50	85.69±0.78	83.66±0.20	50.24±0.76	93.22±0.53	86.07±0.90
AutoGCL	90.27±0.87	89.54±0.80	58.69±1.33	82.73±0.85	82.14±0.66	49.97±0.87	93.16±0.65	86.20±0.42
CCL-Gn	**91.26±0.68**	**90.61±0.34**	**60.21±0.24**	**85.85±1.15**	**84.06±0.65**	51.66±0.22	**94.13±0.46**	**87.28±0.35**

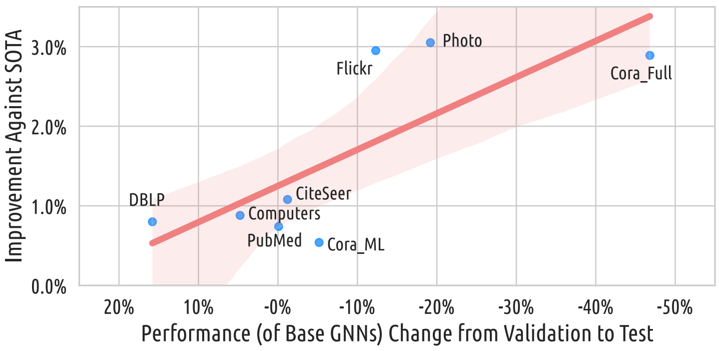


Figure 5: The x-axis is the accuracy change of the base GNNs from the validation set to the test set, *which can imply the degree to which the test set is out-of-distribution*. The y-axis is the improvement in accuracy of CCL-Gn against the best baselines.

more severe the distribution shift in the test set, the greater the advantage of CCL-Gn over SOTA. For example, on Cora_Full, the performance of the base GNNs on the test set decreased by 46.81%, and CCL-Gn improves by 2.89% compared to the best baseline.

Realist Distribution Shift. To further verify the ability of CCL-Gn to enhance the performance of the GNNs in real-world applications, we conducted experiments on two OGB (Hu et al., 2020) datasets, Ogbn-Arxiv and Ogbn-MAG. Note that we only keep citation relations between papers in Ogbn-MAG for the homogeneous setting in this work. These two datasets are at a relatively larger scale and their official data spilts closely reflects real-world application scenarios, with the presence of a distribution shift that is not deliberately designed. Table 4 shows the comparison between CCL-Gn and four promising GCL baselines. CCL-Gn has achieved optimal performance, which validated the trustworthiness of the CCL-Gn outside the laboratory.

Table 4: Comparison with GCL Methods: Node classification accuracy on Ogbn with official splits. The base model is a 2-layer GAT.

	ERM	Grace	GCA	BGRL	AutoGCL	CCL-Gn
Ogbn-Arxiv	68.41±0.75	69.43±0.41	69.35±0.54	68.71±0.60	69.53±0.40	**69.87±0.77**
Ogbn-MAG	32.02±0.75	32.56±0.49	32.39±0.84	31.83±0.48	32.42±0.80	**34.07±1.57**

4.4. Ablation and Sensitivity Analyses

The magnitude of the predictive power of the spurious structures G_v^- compared to the causal structures G_v^+ and the prior proportions of the causally-correlated edges can vary across graphs and tasks. In Equation 9, we introduce two hyperparameters, i.e., α and β, to control these two values, respectively. Here, we analyzed their necessity in CCL-Gn and CCL-Gn 's sensitivity to them (with GAT as the base model). Figure 6 presents the average performances when the hyperparameters of interest were fixed at a number of values. While the results indicated the robustness of CCL-Gn to these hyperparameters in terms of improving model performance, we can find that:

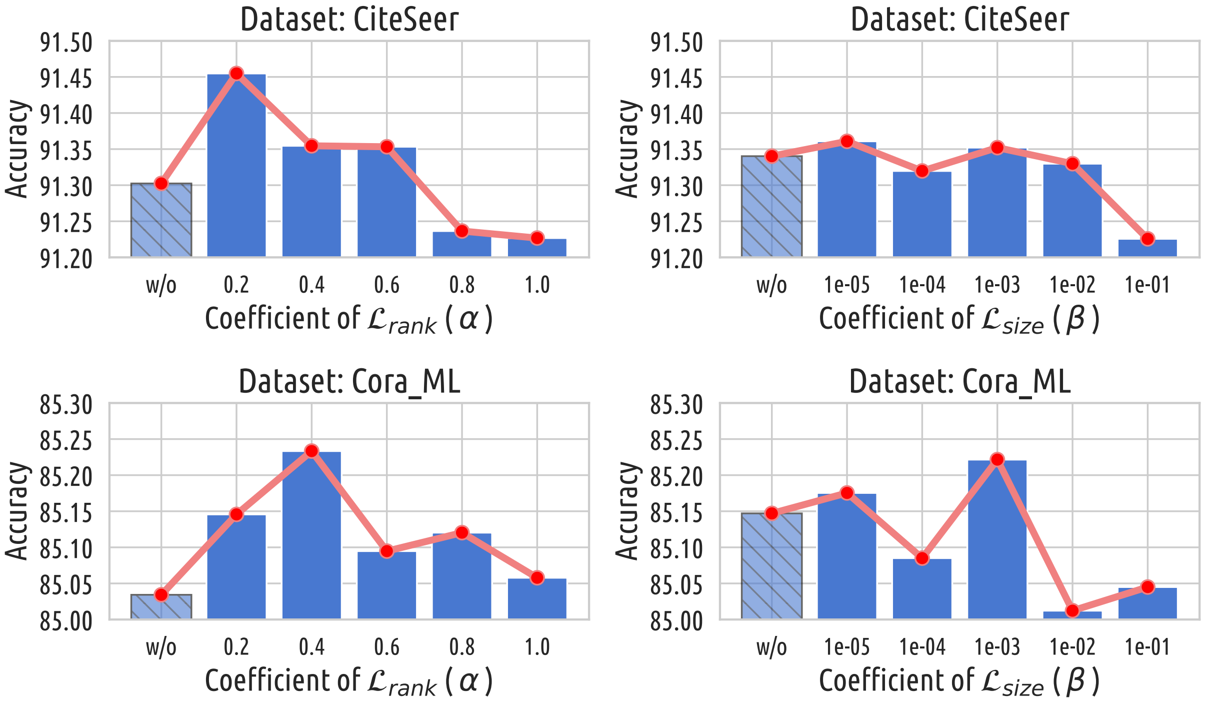


Figure 6: Ablation and sensitivity analyses in terms of the hyperparameter α and β on CiteSeer and Cora_ML with GAT. The first bar labeled "w/o" in the subfigure represents removing the optimization objective, i.e., the ablation study.

1. Both of the two optimization objectives, i.e., $\mathcal{L}_{rank}$ and $\mathcal{L}_{size}$ are important. The performance can be inferior when either of the two optimization objectives is removed. This acknowledges our analysis of the ETSC and the core idea of CCL-Gn.
2. The coefficient of $\mathcal{L}_{rank}$, i.e., α, plays a relatively critical role in CCL-Gn . There is a clear pattern between its value and the model performance: different datasets prefer different optimal values, and the closer its value is to the optimal value on a specific dataset, the better the model performance will be. This pattern reflects the reasonableness of Proposition 3.5.
3. While CCL-Gn is relatively less sensitive to the coefficient of $\mathcal{L}_{size}$, i.e., β, a larger value of β tended to result in inferior performance. This may be due to the fact that when its value is too large, the integrity of the causal structure in the positive view cannot be guaranteed.

Besides, we conduct ablation study on the contrastive objective $\mathcal{L}_{cl}$ in Equation 10 and Equation 12. Experimental results are in Table 5. Generally, the model's performance slightly decreases under the IID setting, while under the OOD setting, the decrease in model performance is relatively more significant. The inferred reasons are as follows. Without contrastive learning, the proposed method CCL-Gn degrades into a state similar to data augmentation. For each ego-graph, the augmentation data includes its causal ego-subgraph and its non-causal ego-subgraph. As all the samples are equivalent during training, the GNNs still focus on all structures. Consequently, under the IID setting, the degraded CCL-Gn still enhances the performance of the GNNs, which is also consistent with the successful experience of general data augmentation under IID settings. However, the decrease in model performance can be more significant under the OOD setting, as data augmentation cannot directly and explicitly address spurious correlations, and training with the augmented non-causal ego-subgraphs may even impair the model's generalizability.

To further understand the impact of generated counterfactual positive and negative views on model performance, we conducted a sensitivity experiment on the temperature parameter τ in the contrastive learning loss in Equation 10. This parameter plays a key role in controlling the penalty strength on the similarity between original ego-graphs and the counterfactual negative views. Specifically, a smaller temperature in the contrastive loss tends to apply stronger penalties, making the influence of generated counterfactual views more significant Wang and Liu (2021). As shown in Figure 7, Overall, performance initially improves as the temperature increases, then declines. This is because a higher temperature reduces the penalty on the similarity between input ego-graphs and counterfactual negative views, thereby lessening the impact of counterfactual views on model training. This trend effectively supports the validity of the generated counterfactual views. However, this parameter should not be set too low, possibly because counterfactual negative views are subgraphs of the input ego-graphs and thus inherently retain some similarity. Forcing this similarity to be too low could lead to model training instability.

Table 5: Ablation study on contrastive learning.

	CiteSeer	Cora_ML	CBASCovariate	CBASConcept	WebKBCovariate	WebKBConcept
Type	IID	IID	OOD	OOD	OOD	OOD
CCL-Gn	91.26±0.68	85.85±1.15	77.14±2.67	85.14±1.06	29.05±6.36	31.19±1.72
w/o $\mathcal{L}_{cl}$	91.18±0.54	85.55±0.77	70.57±1.62	83.85±4.23	16.19±4.91	28.81±0.50

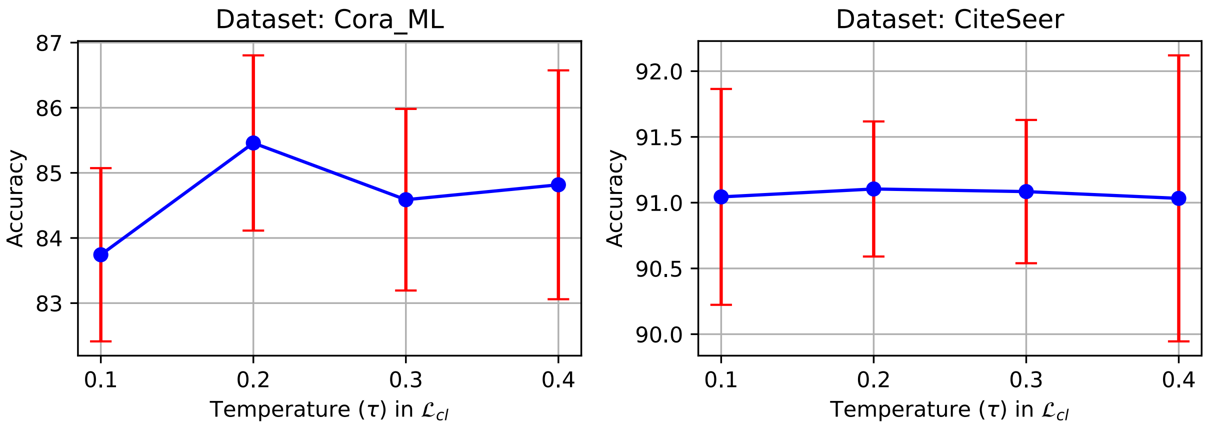


Figure 7: Sensitivity analyses in terms of the hyperparameter τ on CiteSeer and Cora_ML with GAT.

4.5. Discussion on Edge Representations

As aforementioned in Equation 2 and 3, considering the specificity of the ego-graph setting, we additionally add the structural information in the representation of the edge. We display the performance of different edge representations based on the GAT model on CiteSeer and Cora_ML as shown in Figure 8. It's observed that:

1. All the types of edge representation in the counterfactual transformation module can robustly make CCL-Gn work properly. This emphasizes the high-level philosophy of the proposed framework CCL-Gn.

2. Different datasets may prefer different structural information, i.e., ego identifier or distance encoding. Overall, it would not be problematic to include both of them. This highlights the importance of considering the specificity of the edges in the ego-graph compared to the edges in graph tasks.

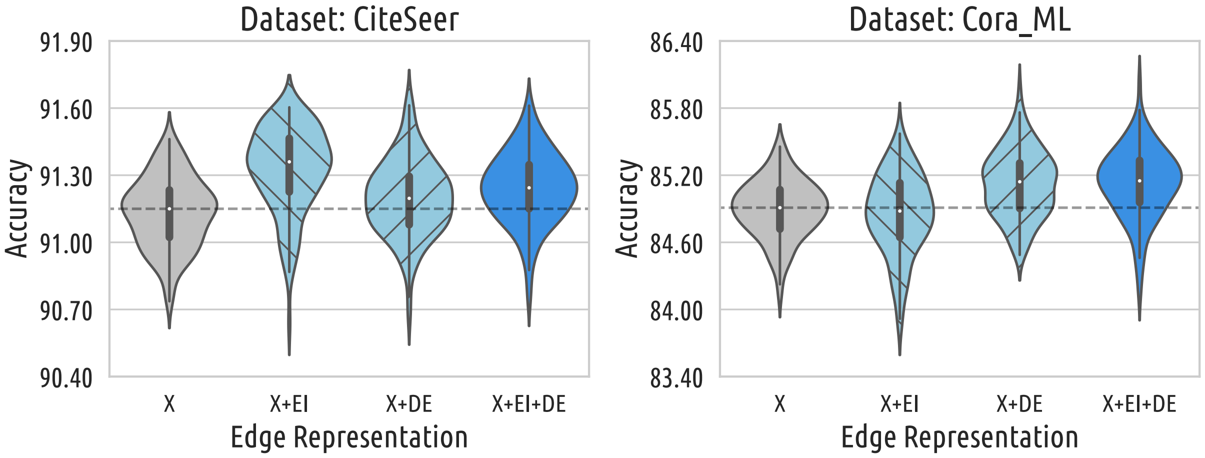


Figure 8: Violin plots of the model performances under different edge representations in CCL-Gn based on the results on CiteSeer and Cora_ML with GAT. "X" represents the raw node feature. "EI" and "DE" represent the ego identifier and distance encoding introduced in Equation 2, respectively.

5. Conclusion

This paper proposes a novel model-agnostic and task-oriented counterfactual contrastive learning framework for graph data in node-level tasks, namely CCL-Gn. It automatically generates optimal counterfactual contrastive view pair via a learnable task-oriented decomposition of the raw ego-graphs. Extensive experiments demonstrated the superiority of CCL-Gn to enhance the performance and the generalizability of a series of typical GNNs without any alteration to the GNN's inference flow.

6. Acknowledgments

This work is supported by the National Natural Science Foundation of China (72134007, 72104212, 62106039), the National Key R&D Program of China (2022ZD0116203), the Natural Science Foundation of Zhejiang Province (LY22G030002), and Alibaba Group through Alibaba Innovative Research Program. We also appreciate the support from the Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

Aldaz, J., Barza, S., Fujii, M., Moslehian, M.S., 2015. Advances in operator cauchy–schwarz inequalities and their reverses. Annals of Functional Analysis 6, 275–295.

An, W., Beauvile, R., Rosche, B., 2022. Causal network analysis. Annual Review of Sociology 48, 23–41.

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

Bojchevski, A., Günnemann, S., 2018. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net. URL: https://openreview.net/forum?id=r1ZdKJ-0W.

Bojchevski, A., Günnemann, S., 2018. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking, in: International Conference on Learning Representations.

Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y., 2020. Counterfactual samples synthesizing for robust visual question answering, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE. pp. 10797–10806. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Chen_Counterfactual_Samples_Synthesizing_for_Robust_Visual_Question_Answering_CVPR_2020_paper.html, doi:10.1109/CVPR42600.2020.01081.

Chen, Y., Bian, Y., Zhou, K., Xie, B., Han, B., Cheng, J., 2023a. Does invariant graph learning via environment augmentation learn invariance?, in: Thirty-seventh Conference on Neural Information Processing Systems.

Chen, Y., Ren, Q., Yong, L., 2023b. Hybrid augmented automated graph contrastive learning. arXiv preprint arXiv:2303.15182 .

Chen, Y., Zhang, Y., Bian, Y., Yang, H., Ma, K., Xie, B., Liu, T., Han, B., Cheng, J., 2022. Learning causally invariant representations for out-of-distribution generalization on graphs, in: Advances in Neural Information Processing Systems.

Choi, S., Jeong, M., Han, H., Hwang, S., 2022. C2L: causally contrastive learning for robust text classification, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, AAAI Press. pp. 10526–10534. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21296.

Dai, E., Jin, W., Liu, H., Wang, S., 2022. Towards robust graph neural networks for noisy graphs with sparse labels, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp. 181–191.

Esser, P., Chennuru Vankadara, L., Ghoshdastidar, D., 2021. Learning theory can (sometimes) explain generalisation in graph neural networks. Advances in Neural Information Processing Systems 34, 27043–27056.

Fan, C., Chen, W., Tian, J., Li, Y., He, H., Jin, Y., 2024. Unlock the potential of counterfactually-augmented data in out-of-distribution generalization. Expert Systems with Applications 238, 122066. URL: https://www.sciencedirect.com/science/article/pii/S095741742302568X, doi:https://doi.org/10.1016/j.eswa.2023.122066.

Fan, S., Wang, X., Mo, Y., Shi, C., Tang, J., 2022a. Debiasing graph neural networks via learning disentangled causal substructure, in: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 24934–24946. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9e47a0bc530cc88b09b7670d2c130a29-Paper-Conference.pdf.

Fan, S., Wang, X., Mo, Y., Shi, C., Tang, J., 2022b. Debiasing graph neural networks via learning disentangled causal substructure, in: NeurIPS.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. Journal of machine learning research 17, 1–35.

Gao, H., Liu, Y., Ji, S., 2021a. Topology-aware graph pooling networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 4512–4518.

Gao, T., Yao, X., Chen, D., 2021b. SimCSE: Simple contrastive learning of sentence embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. pp. 6894–6910. URL: https://aclanthology.org/2021.emnlp-main.552, doi:10.18653/v1/2021.emnlp-main.552.

Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence 2, 665–673.

Grover, A., Leskovec, J., 2016. node2vec: Scalable feature learning for networks, in: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (Eds.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, ACM. pp. 855–864. URL: https://doi.org/10.1145/2939672.2939754, doi:10.1145/2939672.2939754.

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

Hassani, K., Ahmadi, A.H.K., 2020. Contrastive multi-view representation learning on graphs, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, PMLR. pp. 4116–4126. URL: http://proceedings.mlr.press/v119/hassani20a.html.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B., 2020. Momentum contrast for unsupervised visual representation learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE. pp. 9726–9735. URL: https://doi.org/10.1109/CVPR42600.2020.00975, doi:10.1109/CVPR42600.2020.00975.

He, Y., Wang, Z., Cui, P., Zou, H., Zhang, Y., Cui, Q., Jiang, Y., 2022. Causpref: Causal preference learning for out-of-distribution recommendation, in: Proceedings of the ACM Web Conference 2022, Association for Computing Machinery, New York, NY, USA. p. 410–421. URL: https://doi.org/10.1145/3485447.3511969, doi:10.1145/3485447.3511969.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., Leskovec, J., 2020. Open graph benchmark: Datasets for machine learning on graphs. arXiv preprint arXiv:2005.00687 .

Izmailov, P., Kirichenko, P., Gruver, N., Wilson, A.G., 2022. On feature learning in the presence of spurious correlations, in: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 38516–38532. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/fb64a552feda3d981dbe43527a80a07e-Paper-Conference.pdf.

Jackson, M.O., 2005. A survey of network formation models: stability and efficiency. Group formation in economics: Networks, clubs, and coalitions 664, 11–49.

Jang, E., Gu, S., Poole, B., 2017. Categorical reparameterization with gumbel-softmax, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net. URL: https://openreview.net/forum?id=rkE3y85ee.

Jin, W., Zhao, T., Ding, J., Liu, Y., Tang, J., Shah, N., 2022. Empowering graph representation learning with test-time graph transformation, in: The Eleventh International Conference on Learning Representations.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 18661–18673. URL: https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Knyazev, B., Taylor, G.W., Amer, M., 2019. Understanding attention and generalization in graph neural networks. Advances in neural information processing systems 32.

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

Leydesdorff, L., Wagner, C.S., Bornmann, L., 2019. Interdisciplinarity as diversity in citation patterns among journals: Rao-stirling diversity, relative variety, and the gini coefficient. Journal of Informetrics 13, 255–269. URL: https://www.sciencedirect.com/science/article/pii/S1751157718303535, doi:https://doi.org/10.1016/j.joi.2018.12.006.

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

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

Li, H., Zhang, Z., Wang, X., Zhu, W., 2023. Invariant node representation learning under distribution shifts with multiple latent environments. ACM Trans. Inf. Syst. 42. URL: https://doi.org/10.1145/3604427, doi:10.1145/3604427.

Li, P., Wang, Y., Wang, H., Leskovec, J., 2020. Distance encoding: Design provably more powerful neural networks for graph representation learning, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA.

Li, S., Wang, X., Zhang, A., He, X., Chua, T.S., 2022c. Let invariant rationale discovery inspire graph contrastive learning, in: ICML.

Liang, Z., Jiang, W., Hu, H., Zhu, J., 2020. Learning to contrast the counterfactual samples for robust visual question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 3285–3292. URL: https://aclanthology.org/2020.emnlp-main.265, doi:10.18653/v1/2020.emnlp-main.265.

Lin, T., Song, K., Jiang, Z., Kang, Y., Yuan, W., Li, X., Sun, C., Huang, C., Liu, X., 2024. Towards human-like perception: Learning structural causal model in heterogeneous graph. Information Processing & Management 61, 103600. URL: https://www.sciencedirect.com/science/article/pii/S0306457323003370, doi:https://doi.org/10.1016/j.ipm.2023.103600.

Lin, Y., Zhu, S., Tan, L., Cui, P., 2022. Zin: When and how to learn invariance without environment partition? Advances in Neural Information Processing Systems 35, 24529–24542.

Liu, S., Li, T., Feng, Y., Tran, N., Zhao, H., Qiu, Q., Li, P., 2023a. Structural re-weighting improves graph domain adaptation, in: International Conference on Machine Learning, PMLR. pp. 21778–21793.

Liu, Y., Ao, X., Feng, F., Ma, Y., Li, K., Chua, T.S., He, Q., 2023b. Flood: A flexible invariant learning framework for out-of-distribution generalization on graphs, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1548–1558.

Liu, Y., Jin, M., Pan, S., Zhou, C., Zheng, Y., Xia, F., Yu, P., 2022. Graph self-supervised learning: A survey. IEEE Transactions on Knowledge and Data Engineering , 1–1doi:10.1109/TKDE.2022.3172903.

Lloyd, S.P., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28, 129–136. URL: https://doi.org/10.1109/TIT.1982.1056489, doi:10.1109/TIT.1982.1056489.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

Maddison, C.J., Mnih, A., Teh, Y.W., 2017. The concrete distribution: A continuous relaxation of discrete random variables, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net. URL: https://openreview.net/forum?id=S1jE5L5gl.

McAuley, J.J., Leskovec, J., 2012. Image labeling on a network: Using social-network metadata for image classification, in: Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (Eds.), Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV, Springer. pp. 828–841. URL: https://doi.org/10.1007/978-3-642-33765-9_59, doi:10.1007/978-3-642-33765-9_59.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Mo, Y., Wang, X., Fan, S., Shi, C., 2024. Graph contrastive invariant learning from the causal perspective. Proceedings of the AAAI Conference on Artificial Intelligence 38, 8904–8912. URL: https://ojs.aaai.org/index.php/AAAI/article/view/28738, doi:10.1609/aaai.v38i8.28738.

van den Oord, A., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. CoRR abs/1807.03748. URL: http://arxiv.org/abs/1807.03748, arXiv:1807.03748.

Pearl, J., 2009. Causality: Models, Reasoning and Inference. 2nd ed., Cambridge University Press, USA.

Peng, M., Juan, X., Li, Z., 2024. Label-guided graph contrastive learning for semi-supervised node classification. Expert Systems with Applications 239, 122385. URL: https://www.sciencedirect.com/science/article/pii/S0957417423028877, doi:https://doi.org/10.1016/j.eswa.2023.122385.

Pishro-Nik, H., 2014. Introduction to Probability, Statistics, and Random Processes. Kappa Research, LLC. URL: https://books.google.com/books?id=3yq_oQEACAAJ.

Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P., 2019. Distributionally robust neural networks, in: International Conference on Learning Representations.

Sagawa, S., Raghunathan, A., Koh, P.W., Liang, P., 2020. An investigation of why overparameterization exacerbates spurious correlations, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, PMLR. pp. 8346–8356. URL: http://proceedings.mlr.press/v119/sagawa20a.html.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., Mooij, J., 2012. On causal and anticausal learning, in: Proceedings of the 29th International Coference on International Conference on Machine Learning, Omnipress, Madison, WI, USA. p. 459–466.

Schubert, E., Gertz, M., 2018. Numerically stable parallel computation of (co-)variance, in: International Conference, pp. 1–12.

Shchur, O., Mumme, M., Bojchevski, A., Günnemann, S., 2018. Pitfalls of graph neural network evaluation. CoRR abs/1811.05868. URL: http://arxiv.org/abs/1811.05868, arXiv:1811.05868.

Shu, X., Yan, S., Yang, X., Wu, Z., Chen, Z., Lu, Z., 2024. Ascl: Adaptive self-supervised counterfactual learning for robust visual question answering.

Expert Systems with Applications 248, 123125. URL: https://www.sciencedirect.com/science/article/pii/S0957417423036291, doi:https://doi.org/10.1016/j.eswa.2023.123125.

Sun, B., Feng, J., Saenko, K., 2016. Return of frustratingly easy domain adaptation, in: Proceedings of the AAAI conference on artificial intelligence.

Sun, B., Saenko, K., 2016. Deep coral: Correlation alignment for deep domain adaptation, in: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14, Springer. pp. 443–450.

Tan, X., Yong, L., Zhu, S., Qu, C., Qiu, X., Yinghui, X., Cui, P., Qi, Y., 2023. Provably invariant learning without domain information, in: International Conference on Machine Learning, PMLR. pp. 33563–33580.

Thakoor, S., Tallec, C., Azar, M.G., Munos, R., Veličković, P., Valko, M., 2021. Bootstrapped representation learning on graphs, in: ICLR 2021 Workshop on Geometrical and Topological Representation Learning.

Topirceanu, A., Udrescu, M., Marculescu, R., 2018. Weighted betweenness preferential attachment: A new mechanism explaining social network formation and evolution. Scientific reports 8, 10871.

Velickovic, P., Fedus, W., Hamilton, W.L., Liò, P., Bengio, Y., Hjelm, R.D., 2019. Deep graph infomax. ICLR (Poster) 2, 4.

Vigen, T., 2015. Spurious Correlations. Hachette Books. URL: https://books.google.com/books?id=0uDrBQAAQBAJ.

Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. 31, 841.

Wang, F., Liu, H., 2021. Understanding the behaviour of contrastive loss, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2495–2504.

Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A., 2020. Microsoft Academic Graph: When experts are not enough. Quantitative Science Studies 1, 396–413. URL: https://doi.org/10.1162/qss_a_00021, doi:10.1162/qss_a_00021, arXiv:https://direct.mit.edu/qss/article-pdf/1/1/396/1760880/qss_a_00021.pdf.

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

Wei, C., Wang, Y., Bai, B., Ni, K., Brady, D., Fang, L., 2023. Boosting graph contrastive learning via graph contrastive saliency, in: International conference on machine learning, PMLR. pp. 36839–36855.

Wu, Q., Zhang, H., Yan, J., Wipf, D., 2021. Handling distribution shifts on graphs: An invariance perspective, in: International Conference on Learning Representations.

Wu, S., Cao, K., Ribeiro, B., Zou, J., Leskovec, J., 2023. Graphmetro: Mitigating complex distribution shifts in gnns via mixture of aligned experts. arXiv preprint arXiv:2312.04693 .

Wu, Y.X., Wang, X., Zhang, A., He, X., Chua, T.S., 2022. Discovering invariant rationales for graph neural networks. arXiv preprint arXiv:2201.12872 .

Xia, L., Huang, C., Huang, C., Lin, K., Yu, T., Kao, B., 2023. Automated self-supervised learning for recommendation, in: Proceedings of the ACM Web Conference 2023, pp. 992–1002.

Xie, Y., Xu, Z., Zhang, J., Wang, Z., Ji, S., 2022. Self-supervised learning of graph neural networks: A unified review. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–1doi:10.1109/TPAMI.2022.3170559.

Yang, H., Chen, H., Zhang, S., Sun, X., Li, Q., Zhao, X., Xu, G., 2023a. Generating counterfactual hard negative samples for graph contrastive learning, in: Proceedings of the ACM Web Conference 2023, Association for Computing Machinery, New York, NY, USA. p. 621–629. URL: https://doi.org/10.1145/3543507.3583499, doi:10.1145/3543507.3583499.

Yang, M., Fang, Z., Zhang, Y., Du, Y., Liu, F., Ton, J.F., Wang, J., Wang, J., 2023b. Invariant learning via probability of sufficient and necessary causes, in: 2023 Conference on Neural Information Processing Systems.

Yang, Y., Chaudhuri, K., 2022. Understanding rare spurious correlations in neural networks. CoRR abs/2202.05189. URL: https://arxiv.org/abs/2202.05189, arXiv:2202.05189.

Yin, Y., Wang, Q., Huang, S., Xiong, H., Zhang, X., 2022. Autogcl: Automated graph contrastive learning via learnable view generators, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, AAAI Press. pp. 8892–8900. URL: https://ojs.aaai.org/index.php/AAAI/article/view/20871.

Ying, Z., Bourgeois, D., You, J., Zitnik, M., Leskovec, J., 2019. Gnnexplainer: Generating explanations for graph neural networks, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/d80b7040b773199015de6d3b4293c8ff-Paper.pdf.

You, Y., Chen, T., Shen, Y., Wang, Z., 2021. Graph contrastive learning automated, in: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, PMLR. pp. 12121–12132. URL: http://proceedings.mlr.press/v139/you21a.html.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., Shen, Y., 2020. Graph contrastive learning with augmentations, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. URL: https://proceedings.neurips.cc/paper/2020/hash/3fe230348e9a12c13120749e3f9fa4cd-Abstract.html.

Yuster, R., Zwick, U., 2005. Fast sparse matrix multiplication. ACM Trans. Algorithms 1, 2–13. URL: https://doi.org/10.1145/1077464.1077466, doi:10.1145/1077464.1077466.

Zeng, H., Zhou, H., Srivastava, A., Kannan, R., Prasanna, V.K., 2020. Graphsaint: Graph sampling based inductive learning method, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net. URL: https://openreview.net/forum?id=BJe8pkHFwS.

Zhang, H., Wu, Q., Yan, J., Wipf, D., Yu, P.S., 2021a. From canonical correlation analysis to self-supervised graph neural networks. Advances in Neural Information Processing Systems 34, 76–89.

Zhang, S., Yao, D., Zhao, Z., Chua, T., Wu, F., 2021b. Causerec: Counterfactual user sequence synthesis for sequential recommendation, in: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM. pp. 367–377. URL: https://doi.org/10.1145/

3404835.3462908, doi:10.1145/3404835.3462908.

Zhang, Z., Zhao, Z., Lin, Z., He, X., et al., 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. Advances in Neural Information Processing Systems 33, 18123–18134.

Zheng, X., Aragam, B., Ravikumar, P.K., Xing, E.P., 2018. Dags with no tears: Continuous optimization for structure learning, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M., 2020. Graph neural networks: A review of methods and applications. AI open 1, 57–81.

Zhu, Q., Ponomareva, N., Han, J., Perozzi, B., 2021a. Shift-robust gnns: Overcoming the limitations of localized graph training data. Advances in Neural Information Processing Systems 34, 27965–27977.

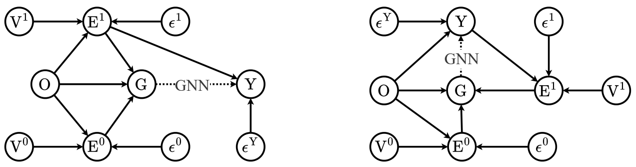
Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., Wang, L., 2020. Deep graph contrastive representation learning. arXiv preprint arXiv:2006.04131 .

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., Wang, L., 2021b. Graph contrastive learning with adaptive augmentation, in: Leskovec, J., Grobelnik, M., Najork, M., Tang, J., Zia, L. (Eds.), WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021, ACM / IW3C2. pp. 2069–2080. URL: https://doi.org/10.1145/3442381.3449802, doi:10.1145/3442381.3449802.

Zhuang, X., Zhang, Q., Ding, K., Bian, Y., Wang, X., Lv, J., Chen, H., Chen, H., 2023. Learning invariant molecular representation in latent discrete space, in: Thirty-seventh Conference on Neural Information Processing Systems.

Appendices

A. Interpretative Examples for Causal Diagrams for 1-hop Ego-Graph Generation Process



(a) Graph structure determines label. (b) Label of ego causes graph structure.

Figure A.9: Causal diagrams for the 1-hop ego-graph generation process. ϵ represents random noise. Solid arrows represent directed causal relationships. The dotted line with text GNN represents that the whole observed graph G is generally used to predict the task label Y of the ego O. **(a)** The ego O can cause two kinds of edges: E^0 (with nodes V^0) and E^1 (with nodes V^1). Only E^1 determines the label Y. **(b)** The ego O itself determines the label Y. Y causes E^1 (with nodes V^1), while O cause E^0 (with nodes V^0).

Here, we provide two interpretative examples as follows.

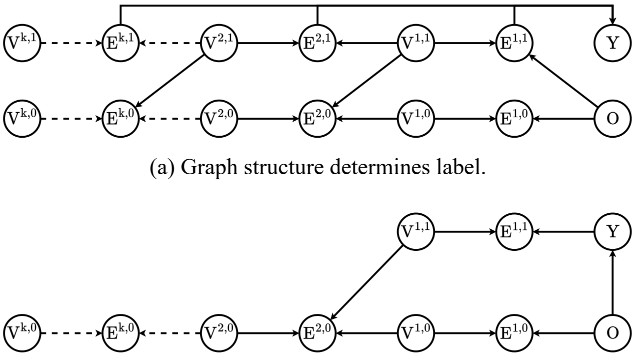
Example 1 (For Figure A.9a)**.** When determining whether a paper is an interdisciplinary work (Y), the decisive factor lies solely in its references (E^1) as the definition of interdisciplinary work is the variety, balance, and disparity of the references Leydesdorff et al. (2019), and the citations of the paper (E^0) are not causal factors. For example, Word2Vec Mikolov et al. (2013) has been utilized in multiple disciplines currently, indicating great interdisciplinarity in the citations (E^0). However, when it was proposed, it was merely a computer science work. In other words, it's possible that a work itself is not interdisciplinary research, but is applied across various disciplines; alternatively, a work itself may be interdisciplinary research but primarily attracts attention from one specific field. Unfortunately, this two variables usually show correlations.

Example 2 (For Figure A.9b)**.** McAuley and Leskovec (2012) proposed an image graph, where an edge exists if two images share common properties, e.g., location, producer, or object. The task is to predict the category of the image (Y), e.g. landscape or city. It's common sense that objects and categories are causally correlated, e.g., the skyscraper and the city. Therefore, E^1 (with V^1) can be connections to images sharing the same objects. And connections to images sharing the same producers should be E^0 (with V^0). However, producers' images are not normally uniformly distributed due to personal interest.

B. Illustration for Causal Diagrams for K-Hop Ego-Graph Generation Process

Let us explain a k-hop ego-graph generation process by a hierarchical traversal (allowing revisiting a node or an edge). The causal diagrams for k-hop ego-graph are shown in Figure B.10.

Figure B.10a. In this case, we assume, when an edge is a causal edge for the task, the edge must have at least one path to the ego. The edges on the path are all considered causal for the task. This is reasonable since the path makes the edge a part of the ego-graph, otherwise, the edge is isolated to the ego-grap and cannot be causally-correlated with the ego. Now if the ego-graph is acyclic, there are three types of edges: (a) edges whose path(s) to the ego only include causal edges (i.e., edges indicated by the arrows in the upper row); (b) edges whose path(s) to the ego include no causal edges (i.e., edges indicated by the arrows in the bottom row); (c) edges whose path(s) to the ego include both causal and non-causal edges (i.e., edges formed by the slanted arrows). The first type can be causally-correlated since they determine the label Y, but the second type can be spuriously-correlated since the association paths from them to the label Y exist fork (confounding) patterns (Pearl, 2009) joined by O. As for the last type, they and label Y can be also confounded by $V^{i,1}$, where $i \in \{0, 1, \ldots, k-1\}$. If the ego-graph has cycles, which means there exists at least one edge that can be included in multiple $E^{i,j}$, where $j \in \{0, 1\}$. This edge is causally-correlated to the label Y if and only if



(a) Graph structure determines label.

(b) Label of ego causes graph structure.

Figure B.10: Causal diagrams for the k-hop ($k > 1$) ego-graph generation process. Random noises are omitted for simplicity.

$\max(j) = 1$, which indicates that this edge is part of the causes of the label Y, otherwise it is spuriously-correlated to the label Y since it can be confounded by O like other typical edges in $\mathrm{E}^{i,0}$.

Figure B.10b. In this case, it's observed that high-order edges have no causal relationship to the label Y. They can be confounded by either O or $\mathrm{V}^{1,1}$.

C. Proof and Simulation for the Proposition 3.5

In this section, we first provide a theoretical proof. Then, we further validate our conclusion through intuitive and visualized simulation results.

C.1. Theoretical Proof

In practice, the function $\phi^{(}G^{-}) = \mathbb{E}[Y|X = G^{+}]$ or $\phi(G^{+}) = \mathbb{E}[Y|X = G^{+}]$ generally have a complicated form. Thus, we might want to use a simpler function ϕ to illustrate the problem. To begin with, we illustrate how to estimate the minimum Mean Squared Error (MSE) of two random variables while using one to regress the other one linearly. Let X and Y be two random variables with finite means and variances. Under a linear function assumption, i.e., $\hat{\mathrm{Y}} = a\mathrm{X} + b$, to estimate the optimal parameters a^* and b^* with the MSE, we have

$$\begin{aligned}
\arg\min_{a,b} \mathrm{MSE}_{\mathrm{X,Y}}(a,b) =& \mathbb{E}[(\mathrm{Y} - \hat{\mathrm{Y}})^2] && \text{(C.1)}\\
=& \mathbb{E}[(\mathrm{Y} - a\mathrm{X} - b)^2] && \text{(C.2)}\\
=& \mathbb{E}[\mathrm{Y}^2 + a^2\mathrm{X}^2 + b^{-}2a\mathrm{YX} - 2b\mathrm{Y} + 2ab\mathrm{X}] && \text{(C.3)}\\
=& \mathbb{E}[\mathrm{Y}^2] + a^2\mathbb{E}[\mathrm{X}^2] + b^2 - 2a\mathbb{E}[\mathrm{YX}] - 2b\mathbb{E}[\mathrm{Y}] + 2ab\mathbb{E}[\mathrm{X}].
\end{aligned}$$

To obtain the optimal a^* and b^*, we can take the derivatives with respect to a and b and set them to zero, i.e.,

$$\begin{aligned}
\mathbb{E}[\mathrm{X}^2] \cdot a + \mathbb{E}[\mathrm{X}] \cdot b - \mathbb{E}[\mathrm{XY}] = 0, && \text{(C.4)}\\
\mathbb{E}[\mathrm{X}] \cdot a + b - \mathbb{E}[\mathrm{Y}] = 0. && \text{(C.5)}
\end{aligned}$$

Using Equation C.5 to represent b in Equation C.4, we have

$$a^* = \frac{\mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]}{\mathbb{E}[X^2] - \mathbb{E}[X]^2} = \frac{\mathrm{Cov}(\mathrm{X}, \mathrm{Y})}{\mathrm{Var}(\mathrm{X})}. \tag{C.6}$$

Then, we can obtain b^* as $\mathbb{E}[\mathrm{Y}] - a^* \cdot \mathbb{E}[\mathrm{X}]$ based on Equation C.5, and $\mathrm{MSE}_{\mathrm{X,Y}}(a^*, b^*)$ is the minimum MSE we can get.

Meanwhile, because for any random variable, the variance of the variable is equal to the mean of the square of the variable minus the square of the mean of the variable (Schubert and Gertz, 2018), we can denote $\mathrm{MSE}_{\mathrm{X,Y}}(a, b)$ as

$$\begin{aligned}\mathrm{MSE}_{\mathrm{X,Y}}(a, b) =& \mathrm{Var}(\mathrm{Y} - a\mathrm{X} - b) + (\mathbb{E}[(\mathrm{Y} - a\mathrm{X} - b)])^2 && \text{(C.7)}\\ =& \mathrm{Var}(\mathrm{Y} - a\mathrm{X} - b) + (\mathbb{E}[\mathrm{Y}] - a \cdot \mathbb{E}[\mathrm{X}] - b)^2. && \text{(C.8)}\end{aligned}$$

Thereby, as Equation C.5 holds, we can have

$$\mathrm{MSE}_{\mathrm{X,Y}}(a^*, b^*) = \mathrm{Var}(\mathrm{Y} - a^*\mathrm{X} - b^*) \tag{C.9}$$

Then, according to (Pishro-Nik, 2014), we can further rewrite $\mathrm{MSE}_{\mathrm{X,Y}}(a^*, b^*)$ as follows:

$$\begin{aligned}\mathrm{MSE}_{\mathrm{X,Y}}(a^*, b^*) =& \mathrm{Var}(\mathrm{Y} - a^*\mathrm{X}) && \text{(C.10)}\\ =& \mathrm{Var}(\mathrm{Y}) + (a^*)^2\mathrm{Var}(\mathrm{X}) - 2a^*\mathrm{Cov}(\mathrm{X}, \mathrm{Y}). && \text{(C.11)}\end{aligned}$$

Using Equation C.6 to denote a in the above equation, we further have

$$\begin{aligned}\mathrm{MSE}_{\mathrm{X,Y}}(a^*, b^*) =& \mathrm{Var}(\mathrm{Y}) + \left(\frac{\mathrm{Cov}(\mathrm{X}, \mathrm{Y})}{\mathrm{Var}(\mathrm{X})}\right)^2 \cdot \mathrm{Var}(\mathrm{X}) - 2 \cdot \left(\frac{\mathrm{Cov}(\mathrm{X}, \mathrm{Y})}{\mathrm{Var}(\mathrm{X})}\right) \cdot \mathrm{Cov}(\mathrm{X}, \mathrm{Y}) && \text{(C.12)}\\ =& \mathrm{Var}(\mathrm{Y}) + \frac{\mathrm{Cov}(\mathrm{X}, \mathrm{Y})^2}{\mathrm{Var}(\mathrm{X})} - 2 \cdot \frac{\mathrm{Cov}(\mathrm{X}, \mathrm{Y})^2}{\mathrm{Var}(\mathrm{X})} && \text{(C.13)}\\ =& \mathrm{Var}(\mathrm{Y}) - \frac{\mathrm{Cov}(\mathrm{X}, \mathrm{Y})^2}{\mathrm{Var}(\mathrm{X})} && \text{(C.14)}\end{aligned}$$

Based on this finding, we investigate the minimum MSE of regressing Y using G^+ and G^- under the two causal diagrams.

C.1.1. Case A: Graph structure determines the label.

Firstly, according to the causal diagrams, we assume

$$\begin{aligned}& O \sim \mathcal{N}(\mu^O, \sigma^O), \quad \sigma^O \neq 0 && \text{(C.15)}\\ & G^+ = p^+ \cdot O + q^+ + \epsilon^+, \quad \epsilon^+ \sim \mathcal{N}(\mu^+, \sigma^+), \quad p^+ \neq 0, \quad \sigma^+ \neq 0 && \text{(C.16)}\\ & G^- = p^- \cdot O + q^- + \epsilon^-, \quad \epsilon^- \sim \mathcal{N}(\mu^-, \sigma^-), \quad p^- \neq 0, \quad \sigma^- \neq 0 && \text{(C.17)}\\ & Y = p^Y \cdot G^+ + q^Y + \epsilon^Y, \quad \epsilon^Y \sim \mathcal{N}(\mu^Y, \sigma^Y), \quad p^Y \neq 0, \quad \sigma^Y \neq 0 && \text{(C.18)}\end{aligned}$$

where O, ϵ^+, ϵ^-, ϵ^Y are random noises and follow a normal distribution. Note that ϵ^+, ϵ^- are independent of O and ϵ^Y are independent of G^+. Note that $p \cdot O + q$ with $O \sim \mathcal{N}(\mu, \sigma)$ can be rewritten as $p^{'} \cdot O^{'} + q^{'}$ where $p^{'} = p\sigma$, $q^{'} = q + p\mu$ and $O^{'} \sim \mathcal{N}(0, 1)$. Similarly, $q + \epsilon$ with $\epsilon \sim \mathcal{N}(\mu, \sigma)$ can be rewritten as $q^{'} + \epsilon^{'}$ where $q^{'} = q + \mu$ and $\epsilon^{'} \sim \mathcal{N}(0, \sigma)$. Therefore, without loss of generalizability, we assume that μ^O, μ^+, μ^-, μ^Y are equal to zero and σ^O is equal to one.

Then to prove $\mathrm{MSE}^*_{G^+,Y} < \mathrm{MSE}^*_{G^-,Y}$, we can prove $\frac{\mathrm{Cov}(\mathrm{G}^+,\mathrm{Y})^2}{\mathrm{Var}(\mathrm{G}^+)} - \frac{\mathrm{Cov}(\mathrm{G}^-,\mathrm{Y})^2}{\mathrm{Var}(\mathrm{G}^-)} > 0$ based on Equation C.14. $\frac{\mathrm{Cov}(\mathrm{G}^+,\mathrm{Y})^2}{\mathrm{Var}(\mathrm{G}^+)}$,

$\frac{\mathrm{Cov(G^{-},Y)^2}}{\mathrm{Var(G^{-})}}$ are computed as follows:

$$
\begin{aligned}
\frac{\mathrm{Cov(G^{+},Y)^2}}{\mathrm{Var(G^{+})}} &= \frac{(\mathbb{E}[G^{+}\cdot Y]-\mathbb{E}[G^{+}]\cdot\mathbb{E}[Y])^2}{\mathrm{Var}(p^{+}\cdot O)+\mathrm{Var}(\epsilon^{+})} && \text{(C.19)}\\
&= \frac{\Big(\mathbb{E}[G^{+}\cdot(p^{Y}\cdot G^{+}+q^{Y}+\epsilon^{Y})]-q^{+}\cdot(p^{Y}q^{+}+q^{Y})\Big)^2}{(p^{+})^2+(\sigma^{+})^2} && \text{(C.20)}\\
&= \frac{\Big(p^{Y}\cdot\mathbb{E}[(G^{+})^2]+q^{Y}\cdot\mathbb{E}[G^{+}]-q^{+}\cdot(p^{Y}q^{+}+q^{Y})\Big)^2}{(p^{+})^2+(\sigma^{+})^2} && \text{(C.21)}\\
&= \frac{\Big(p^{Y}\cdot(\mathrm{Var}(G^{+})+\mathbb{E}[(G^{+})]^2)+q^{Y}\cdot\mathbb{E}[G^{+}]-q^{+}\cdot(p^{Y}q^{+}+q^{Y})\Big)^2}{(p^{+})^2+(\sigma^{+})^2} && \text{(C.22)}\\
&= \frac{\Big(p^{Y}\cdot((p^{+})^2+(\sigma^{+})^2+(q^{+})^2)+q^{Y}\cdot q^{+}-q^{+}\cdot(p^{Y}q^{+}+q^{Y})\Big)^2}{(p^{+})^2+(\sigma^{+})^2} && \text{(C.23)}\\
&= \frac{(p^{Y})^2\cdot\Big((p^{+})^2+(\sigma^{+})^2\Big)^2}{(p^{+})^2+(\sigma^{+})^2} && \text{(C.24)}\\
&= (p^{Y})^2\cdot\Big((p^{+})^2+(\sigma^{+})^2\Big) && \text{(C.25)}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{Cov(G^{-},Y)^2}}{\mathrm{Var(G^{-})}} &= \frac{(\mathbb{E}[G^{-}\cdot Y]-\mathbb{E}[G^{-}]\cdot\mathbb{E}[Y])^2}{\mathrm{Var}(p^{-}\cdot O)+\mathrm{Var}(\epsilon^{-})} && \text{(C.26)}\\
&= \frac{\Big(\mathbb{E}[G^{-}\cdot\Big(p^{Y}\cdot G^{+}+q^{Y}+\epsilon^{Y}\Big)]-q^{-}\cdot\Big(p^{Y}q^{+}+q^{Y}\Big)\Big)^2}{(p^{-})^2+(\sigma^{-})^2} && \text{(C.27)}\\
&= \frac{\Big(p^{Y}\mathbb{E}[G^{-}\cdot G^{+}]+q^{Y}\mathbb{E}[G^{-}]-q^{-}\cdot\Big(p^{Y}q^{+}+q^{Y}\Big)\Big)^2}{(p^{-})^2+(\sigma^{-})^2} && \text{(C.28)}\\
&= \frac{\Big(p^{Y}\mathbb{E}[(p^{-}\cdot O+q^{-}+\epsilon^{-})\cdot G^{+}]+q^{Y}\cdot q^{-}-q^{-}\cdot\Big(p^{Y}q^{+}+q^{Y}\Big)\Big)^2}{(p^{-})^2+(\sigma^{-})^2} && \text{(C.29)}\\
&= \frac{\Big(p^{Y}\cdot p^{-}\mathbb{E}[O\cdot G^{+}]+p^{Y}\cdot q^{-}\mathbb{G}^{+}+q^{Y}\cdot q^{-}-q^{-}\cdot\Big(p^{Y}q^{+}+q^{Y}\Big)\Big)^2}{(p^{-})^2+(\sigma^{-})^2} && \text{(C.30)}\\
&= \frac{\Big(p^{Y}\cdot p^{-}\cdot\mathbb{E}[O\cdot G^{+}]+p^{Y}\cdot q^{-}\cdot\mathbb{E}[G^{+}]+q^{Y}\cdot q^{-}-q^{-}\cdot\Big(p^{Y}q^{+}+q^{Y}\Big)\Big)^2}{(p^{-})^2+(\sigma^{-})^2} && \text{(C.31)}\\
&= \frac{\Big(p^{Y}\cdot p^{-}\cdot\mathbb{E}[O\cdot(p^{+}\cdot O+q^{+}+\epsilon^{+})]+p^{Y}\cdot q^{-}\cdot q^{+}+q^{Y}\cdot q^{-}-q^{-}\cdot\Big(p^{Y}q^{+}+q^{Y}\Big)\Big)^2}{(p^{-})^2+(\sigma^{-})^2} && \text{(C.32)}\\
&= \frac{\Big(p^{Y}\cdot p^{-}\cdot\Big(p^{+}\cdot\mathbb{E}[O^2]+q^{+}\cdot\mathbb{E}[O]\Big)+p^{Y}\cdot q^{-}\cdot q^{+}+q^{Y}\cdot q^{-}-q^{-}\cdot\Big(p^{Y}q^{+}+q^{Y}\Big)\Big)^2}{(p^{-})^2+(\sigma^{-})^2} && \text{(C.33)}\\
&= \frac{\Big(p^{Y}\cdot p^{-}\cdot\Big(p^{+}\cdot(\mathrm{Var}(O)-\mathbb{E}[O]^2)\Big)+p^{Y}\cdot q^{-}\cdot q^{+}+q^{Y}\cdot q^{-}-q^{-}\cdot\Big(p^{Y}q^{+}+q^{Y}\Big)\Big)^2}{(p^{-})^2+(\sigma^{-})^2} && \text{(C.34)}\\
&= \frac{\Big(p^{Y}\cdot p^{-}\cdot p^{+}+p^{Y}\cdot q^{-}\cdot q^{+}+q^{Y}\cdot q^{-}-q^{-}\cdot\Big(p^{Y}q^{+}+q^{Y}\Big)\Big)^2}{(p^{-})^2+(\sigma^{-})^2} && \text{(C.35)}\\
&= \frac{\Big(p^{Y}\cdot p^{-}\cdot p^{+}\Big)^2}{(p^{-})^2+(\sigma^{-})^2} && \text{(C.36)}
\end{aligned}
$$

Therefore, we can prove $\frac{\mathrm{Cov(G^+,Y)^2}}{\mathrm{Var(G^+)}} - \frac{\mathrm{Cov(G^-,Y)^2}}{\mathrm{Var(G^-)}} > 0$ as

$$
\begin{aligned}
&(p^Y)^2 \cdot \Big((p^+)^2 + (\sigma^+)^2\Big) - \frac{\Big(p^Y \cdot p^- \cdot p^+\Big)^2}{(p^-)^2 + (\sigma^-)^2} > 0 && \text{(C.37)}\\
\iff &\Big((p^+)^2 + (\sigma^+)^2\Big) - \frac{(p^- \cdot p^+)^2}{(p^-)^2 + (\sigma^-)^2} > 0 && \text{(C.38)}\\
\iff &\Big((p^+)^2 + (\sigma^+)^2\Big) \cdot \Big((p^-)^2 + (\sigma^-)^2\Big) - (p^- \cdot p^+)^2 > 0 && \text{(C.39)}\\
\iff &(p^+)^2 \cdot (\sigma^-)^2 + (\sigma^+)^2 \cdot (p^-)^2 + (\sigma^+)^2 \cdot (\sigma^-)^2 > 0. && \text{(C.40)}
\end{aligned}
$$

Because p^+, p^-, σ^+, and σ^- are not equal to 0, the equation holds true, which proves the proposition.

C.1.2. Case B: Label of ego causes graph structure.

Similarly, we first assume

$$
\begin{aligned}
O &\sim \mathcal{N}(\mu^O, \sigma^O), \;\; \sigma^O \neq 0 && \text{(C.41)}\\
G^- &= p^- \cdot O + q^- + \epsilon^-, \;\; \epsilon^- \sim \mathcal{N}(\mu^-, \sigma^-), \;\; p^- \neq 0, \;\; \sigma^- \neq 0 && \text{(C.42)}\\
Y &= p^Y \cdot O + q^Y + \epsilon^Y, \;\; \epsilon^Y \sim \mathcal{N}(\mu^Y, \sigma^Y), \;\; p^Y \neq 0, \;\; \sigma^Y \neq 0 && \text{(C.43)}\\
G^+ &= p^+ \cdot Y + q^+ + \epsilon^+, \;\; \epsilon^+ \sim \mathcal{N}(\mu^+, \sigma^+), \;\; p^+ \neq 0, \;\; \sigma^+ \neq 0 && \text{(C.44)}
\end{aligned}
$$

where O, ϵ^+, ϵ^-, ϵ^Y are random noises and follow a normal distribution. ϵ^+, ϵ^- are independent of O and ϵ^Y are independent of G^+. without loss of generalizability, μ^O, μ^+, μ^-, μ^Y are set to zero and σ^O is set to one.

$\frac{\mathrm{Cov(G^+,Y)^2}}{\mathrm{Var(G^+)}}$, $\frac{\mathrm{Cov(G^-,Y)^2}}{\mathrm{Var(G^-)}}$ are computed as follows:

$$
\begin{aligned}
\frac{\mathrm{Cov(G^+, Y)^2}}{\mathrm{Var(G^+)}} &= \frac{(\mathbb{E}[G^+ \cdot Y] - \mathbb{E}[G^+] \cdot \mathbb{E}[Y])^2}{\mathrm{Var}(p^+ \cdot \mathrm{Y}) + \mathrm{Var}(\epsilon^+)} && \text{(C.45)}\\
&= \frac{\Big(\mathbb{E}[(p^+Y + q^+ + \epsilon^+) \cdot Y] - \mathbb{E}[(p^+Y + q^+ + \epsilon^+)] \cdot \mathbb{E}[(p^YO + q^Y + \epsilon^Y)]\Big)^2}{(p^+)^2 \cdot \mathrm{Var(Y)} + (\sigma^+)^2} && \text{(C.46)}\\
&= \frac{\Big(p^+ \cdot \mathbb{E}[(Y)^2] + q^+ \cdot \mathbb{E}[Y] - (p^+\mathbb{E}[Y] + q^+) \cdot q^Y\Big)^2}{(p^+)^2 \cdot \mathrm{Var}(p^\mathrm{Y} \cdot O + \epsilon^\mathrm{Y}) + (\sigma^+)^2} && \text{(C.47)}\\
&= \frac{\Big(p^+ \cdot (\mathrm{Var}(Y) + \mathbb{E}[(Y)]^2) + q^+ \cdot q^Y - (p^+q^Y + q^+) \cdot q^Y\Big)^2}{(p^+)^2 \cdot ((p^Y)^2 + (\sigma^Y)^2) + (\sigma^+)^2} && \text{(C.48)}\\
&= \frac{\Big(p^+ \cdot ((p^Y)^2 + (\sigma^Y)^2 + (q^Y)^2) + q^+ \cdot q^Y - (p^+q^Y + q^+) \cdot q^Y\Big)^2}{(p^+)^2 \cdot ((p^Y)^2 + (\sigma^Y)^2) + (\sigma^+)^2} && \text{(C.49)}\\
&= \frac{(p^+)^2 \cdot \Big((p^Y)^2 + (\sigma^Y)^2\Big)^2}{(p^+)^2 \cdot ((p^Y)^2 + (\sigma^Y)^2) + (\sigma^+)^2} && \text{(C.50)}
\end{aligned}
$$

$$\frac{\mathrm{Cov(G^{-}, Y)}^2}{\mathrm{Var(G^{-})}} = \frac{(\mathbb{E}[G^{-} \cdot Y] - \mathbb{E}[G^{-}] \cdot \mathbb{E}[Y])^2}{\mathrm{Var}(p^{-} \cdot O) + \mathrm{Var}(\epsilon^{-})} \tag{C.51}$$

$$= \frac{\left(\mathbb{E}[(p^{-}O + q - +\epsilon^{-}) \cdot (p^{Y}O + q^{Y} + \epsilon^{Y})] - q^{-} \cdot q^{Y}\right)^2}{(p^{-})^2 + (\sigma^{-})^2} \tag{C.52}$$

$$= \frac{\left(p^{-} \cdot p^{Y}\mathbb{E}[O^2] + q - \cdot q^{Y} - q^{-} \cdot q^{Y}\right)^2}{(p^{-})^2 + (\sigma^{-})^2} \tag{C.53}$$

$$= \frac{\left(p^{-} \cdot p^{Y} \cdot (\mathrm{Var}(O) + \mathrm{E}[O]^2)\right)^2}{(p^{-})^2 + (\sigma^{-})^2} \tag{C.54}$$

$$= \frac{\left(p^{-} \cdot p^{Y}\right)^2}{(p^{-})^2 + (\sigma^{-})^2} \tag{C.55}$$

Therefore, we can prove $\frac{\mathrm{Cov(G^{+},Y)}^2}{\mathrm{Var(G^{+})}} - \frac{\mathrm{Cov(G^{-},Y)}^2}{\mathrm{Var(G^{-})}} > 0$ as

$$\frac{(p^{+})^2 \cdot \left((p^{Y})^2 + (\sigma^{Y})^2\right)^2}{(p^{+})^2 \cdot ((p^{Y})^2 + (\sigma^{Y})^2) + (\sigma^{+})^2} - \frac{\left(p^{-} \cdot p^{Y}\right)^2}{(p^{-})^2 + (\sigma^{-})^2} > 0 \tag{C.56}$$

$$\iff (p^{-}p^{+})^2(\sigma^{Y})^4 + (p^{-}p^{+}p^{Y}\sigma^{Y})^2 + (p^{+}\sigma^{-})^2(p^{Y})^4 + (p^{+}\sigma^{-})^2(\sigma^{Y})^4 + 2 \cdot (p^{+}p^{Y}\sigma^{Y}\sigma^{-})^2 - (\sigma^{+}p^{-}p^{Y})^2 > 0 \tag{C.57}$$

$$\Longleftarrow \frac{((p^{-}p^{+})(\sigma^{Y})^2 + (p^{-}p^{+}p^{Y}\sigma^{Y}) + (p^{+}\sigma^{-})(p^{Y})^2 + (p^{+}\sigma^{-})(\sigma^{Y})^2 + 2 \cdot (p^{+}p^{Y}\sigma^{Y}\sigma^{-}))^2}{6} - (\sigma^{+}p^{-}p^{Y})^2 > 0 \tag{C.58}$$

$$\text{which is transformed based on Cauchy–Schwarz inequality (Aldaz et al., 2015)} \tag{C.59}$$

$$\iff \frac{(p^{-}p^{+})(\sigma^{Y})^2 + (p^{-}p^{+}p^{Y}\sigma^{Y}) + (p^{+}\sigma^{-})(p^{Y})^2 + (p^{+}\sigma^{-})(\sigma^{Y})^2 + 2 \cdot (p^{+}p^{Y}\sigma^{Y}\sigma^{-})}{\sqrt{6}} - \sigma^{+}p^{-}p^{Y} > 0 \tag{C.60}$$

$$\iff \frac{\frac{(\sigma^{Y})^2}{p^{Y}} + \sigma^{Y} + \frac{\sigma^{-}p^{Y}}{p^{-}} + \frac{\sigma^{-}(\sigma^{Y})^2}{p^{-}p^{Y}} + 2 \cdot \frac{\sigma^{Y}\sigma^{-}}{p^{-}}}{\sqrt{6}} - \frac{\sigma^{+}}{p^{+}} > 0 \tag{C.61}$$

$$\iff \frac{c + 1 + \frac{b}{c} + bc + 2b}{\sqrt{6}} - \sqrt{\frac{1}{c^2} + 1} \cdot a > 0, \tag{C.62}$$

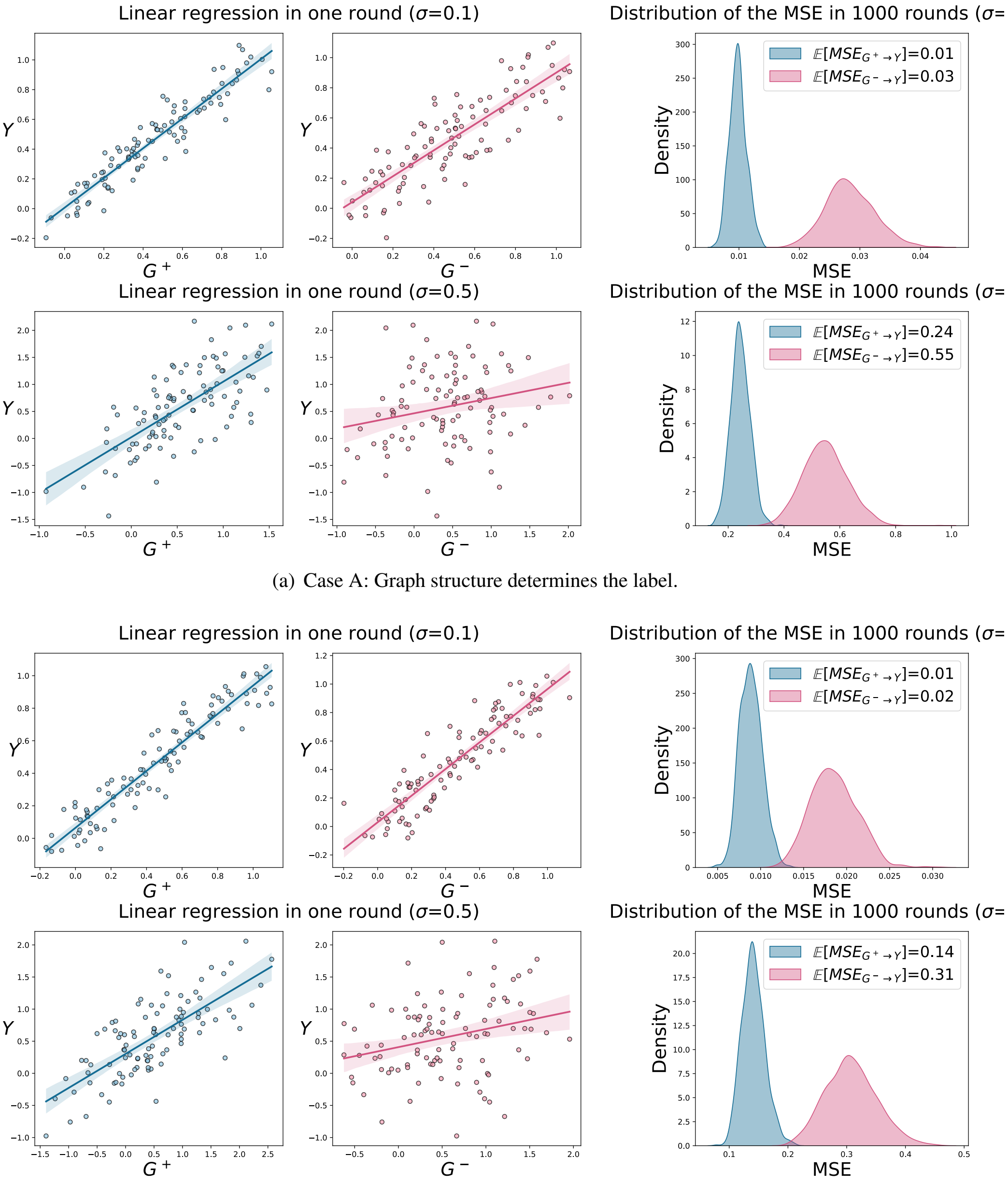
where we denote $a^2 = \frac{(\sigma^{+})^2}{(p^{+})^2 \cdot ((p^{Y})^2 + (\sigma^{Y})^2)}$, $b^2 = \frac{(\sigma^{-})^2}{(p^{-})^2}$, $c^2 = \frac{(\sigma^{Y})^2}{(p^{Y})^2}$. Note that a^2,b^2, and c^2 indicate the proportion of the variance of the random noise (i.e., ϵ^{+}, ϵ^{-} and ϵ^{Y}) in the variance of the variables themselves (i.e., G^{+}, G^{-} and Y). Now that we have Equation C.62 holds true when

$$a^2 < \frac{(c(c+1)(c+2))^2}{6(1+c^2)} \iff k^2 < \frac{((c+1)(c+2))^2}{6(1+c^2)}, \text{ denoting } a^2 \text{ as } (kc)^2 \tag{C.63}$$

This is a monotonically increasing function on the interval [0, 1]. While $c^2 \gtrsim 0.023$, we have $k \gtrsim 1$. In other words, for Y and G^{-}, if the proportion of the variance of the random noise (i.e., ϵ^{Y} and ϵ^{-}) in the variance of the variables themselves (i.e., Y and G^{-}) exceeds 2.25% ($\frac{0.023}{1+0.023}$), the Equation C.62 holds true as long as the proportion of the variance of random noise ϵ^{+} in the variance of the variable G^{+} does not exceed the proportion of the variance of the random noise in the variance of the variables Y and G^{-}. In real-world applications, the proportion of noise is generally not this small. Therefore, the proposition can hold true in most real-life scenarios.

C.2. Simulation Validation

Building on the theoretical analysis above, we provide a visual simulation experiment in Figure C.11 for the two cases, which can clearly demonstrate that predicting Y using G^{-} leads to larger errors compared to predicting Y using G^{+}. This can further help us intuitively understand the proposition 3.5.



(a) Case A: Graph structure determines the label.

(b) Case B: Label of ego causes graph structure.

Figure C.11: Simulation for the Proposition 3.5. For simplicity, we set $\sigma^+ = \sigma^- = \sigma^Y = \sigma$ and $p^+ = p^- = p^Y = 1$.

D. Discussion on the Form of the Objective Function $\mathcal{L}_{rank}$

Under Proposition 3.5, the relationship between G_v^+ and G_v^- can be established by a rank loss $\mathcal{L}_{rank}(\varphi \mid \hat{\phi}^+, \hat{\phi}^-)$ that enforces $\hat{\phi}^-(G_v^-; \hat{\Theta}^-)$ to have greater empirical loss (than $\hat{\phi}^+$) with sufficient predictive power.

Here we illustrate the reason why choose the fraction form in $\mathcal{L}_{rank}$. Still using the setting in the theoretical proof and simulation as an example, to achieve the goal of ranking, the marginal rank loss is a common choice, i.e., ensuring $\text{MSE}^*_{G^-,Y} - \text{MSE}^*_{G^+,Y} > \delta$. However, given our intermediate results from the proof, we find that the difference between the two empirical errors can vary when the magnitude and proportion of random noise (the σ) change. Moreover, this variation can be of a significant magnitude, e.g., in the Figure C.11, when σ changes from 0.1 to 0.5, the difference can change from 0.02/0.02 to 0.31/0.17 (more than ten times). In fact, it can be infinitely large. For example, in case A, this difference is $(p^+)^2 \cdot (\sigma^-)^2 + (\sigma^+)^2 \cdot (p^-)^2 + (\sigma^+)^2 \cdot (\sigma^-)^2$, and it increases as p^+, p^-, σ^+ and σ^- increase without limitation. Therefore, it's hard for us to choose a suitable δ.

However, we find that using a division to represent this magnitude relationship can confine the value of this objective function to between zero and one. Denoting $\frac{\mathrm{Cov(G^+,Y)^2}}{\mathrm{Var(G^+)}}$, $\frac{\mathrm{Cov(G^-,Y)^2}}{\mathrm{Var(G^-)}}$ as $h_{G^+,Y}$, $h_{G^-,Y}$, respectively. For case A, we have

$$\frac{\mathrm{Var(Y)} - h_{G^+,Y}}{\mathrm{Var(Y)} - h_{G^-,Y}} = 1 - \frac{h_{G^+,Y} - h_{G^-,Y}}{\mathrm{Var(Y)} - h_{G^-,Y}} \tag{D.1}$$

$$= 1 - \frac{(p^Y)^2 \cdot \left((p^+)^2 + (\sigma^+)^2\right) - \frac{\left(p^Y \cdot p^- \cdot p^+\right)^2}{(p^-)^2+(\sigma^-)^2}}{(p^Y)^2\left((p^+)^2 + (\sigma^+)^2\right) + (\sigma^Y)^2 - \frac{\left(p^Y \cdot p^- \cdot p^+\right)^2}{(p^-)^2+(\sigma^-)^2}} \tag{D.2}$$

$$= 1 - \frac{(p^Y)^2 \cdot \left((p^+)^2 + (\sigma^+)^2\right) - \frac{\left(p^Y \cdot p^- \cdot p^+\right)^2}{(p^-)^2+(\sigma^-)^2} + (\sigma^Y)^2 - (\sigma^Y)^2}{(p^Y)^2\left((p^+)^2 + (\sigma^+)^2\right) + (\sigma^Y)^2 - \frac{\left(p^Y \cdot p^- \cdot p^+\right)^2}{(p^-)^2+(\sigma^-)^2}} \tag{D.3}$$

$$= \frac{(\sigma^Y)^2}{(p^Y)^2\left((p^+)^2 + (\sigma^+)^2\right) + (\sigma^Y)^2 - \frac{\left(p^Y \cdot p^- \cdot p^+\right)^2}{(p^-)^2+(\sigma^-)^2}} \tag{D.4}$$

$$= \frac{\left((p^-)^2 + (\sigma^-)^2\right) \cdot (\sigma^Y)^2}{(p^Y)^2\left((p^-)^2 + (\sigma^-)^2\right) \cdot \left((p^+)^2 + (\sigma^+)^2\right) + \left((p^-)^2 + (\sigma^-)^2\right) \cdot (\sigma^Y)^2 - \left(p^Y \cdot p^- \cdot p^+\right)^2} \tag{D.5}$$

$$= \frac{\left((p^-)^2 + (\sigma^-)^2\right) \cdot (\sigma^Y)^2}{(p^Y)^2\left((p^+)^2 \cdot (\sigma^-)^2 + (\sigma^+)^2 \cdot (p^-)^2 + (\sigma^+)^2 \cdot (\sigma^-)^2\right) + \left((p^-)^2 + (\sigma^-)^2\right) \cdot (\sigma^Y)^2} \tag{D.6}$$

$$= \frac{1}{\frac{(p^Y)^2\left((p^+)^2 \cdot (\sigma^-)^2 + (\sigma^+)^2 \cdot (p^-)^2 + (\sigma^+)^2 \cdot (\sigma^-)^2\right)}{\left((p^-)^2+(\sigma^-)^2\right) \cdot (\sigma^Y)^2} + 1}, \tag{D.7}$$

which proves that in case A the division is bounded from 0 to 1. Similarly, For case B, we will have

$$\frac{\mathrm{Var(Y)} - h_{G^+,Y}}{\mathrm{Var(Y)} - h_{G^-,Y}} = 1 - \frac{(b^2 + b^2c^2) \cdot (\sigma^+)^2}{a^2 \cdot (b^2 + b^2c^2 + c^2)}, \tag{D.8}$$

which proves that in case B the division is bounded from 0 to 1. Therefore, without choosing a hyperparameter δ, we directly optimize this objective.

E. Reparameterization Trick

Denoting $\frac{1}{m} \sum_{i=1}^{m} \mathrm{MLP}_i \left(\boldsymbol{e}_{u\varsigma|G_v}; \Lambda_i\right)$ in Equation 4 by $s_{u\varsigma|G_v}$, in the training stage, the probability of the edge (u, ς) being part of E_v^+ is given by

$$\theta_{u\varsigma|G_v}^{train} = S\left(\frac{\log(\epsilon) - \log(1-\epsilon) + s_{u\varsigma|G_v}}{\lambda}\right),$$

where $S(x) = \frac{1}{1+e^{-x}}$ and $\epsilon \sim \mathcal{U}(0,\ 1)$ is an independent random variable that obeys a standard uniform distribution. λ is the temperature parameter to control the approximation. When $\lambda \to 0$, $\theta_{u\varsigma|G_v}^{train}$ is binarized with

$$\lim_{\tau \to 0} P\left(\theta_{u\varsigma|G_v}^{train} = 1\right) = \frac{\exp\left(s_{u\varsigma|G_v}\right)}{1 + \exp\left(s_{u\varsigma|G_v}\right)}.$$

F. Dataset Details

F.1. The Synthetic Dataset

We assume that nodes have two shapes, namely square and circular, and two colors, namely red and blue. These two characteristics are latent variables and assign each node an eight-dimensional feature vector, with four dimensions generated by shape and four dimensions generated by color. Note that the eight dimensions are mutually independent of each other. In detail, if the node's shape is square, then the first four features are independently and repeatedly sampled from a normal distribution $\mathcal{N}(-1, 1)$; if the node's shape is circular, the first four features are independently and repeatedly sampled from a normal distribution $\mathcal{N}(1, 1)$; if the node's color is red, then the last four features are independently and repeatedly sampled from a normal distribution $\mathcal{N}(1, 1)$; if the node's color is blue, the last four features are independently and repeatedly sampled from a normal distribution $\mathcal{N}(-1, 1)$. Each node is randomly connected to 8 to 12 neighbors while ensuring approximately a 60% probability of sharing the same color or shape with its neighbors. Next, we sample 100 nodes for each combination of different colors and shapes, and divide them into training, validation, and test sets in a ratio of 4:3:3. The task is to infer the shape and color of each node based on the 1-hop ego-graph of the node. The experimental results of these two tasks are shown in Table F.6

Table F.6: Performance of the color and shape prediction tasks on the synthetic graph with GAT as base model.

	Color	Shape
ERM	67.16±4.55	83.83±3.56
CCL-Gn	71.17±3.21	85.50±2.25

F.2. Datasets in the Comparison with Graph OOD Methods

GOOD-Cora is a citation network adapted from the full Cora dataset Bojchevski and Günnemann (2018). The input is a small-scale citation network graph, in which nodes represent scientific publications and edges are citation links. The task is a 70-class classification of publication types. GOOD Gui et al. (2022) generates splits based on two domain selections, namely, word and degree. The first one is the word diversity defined by the selected word count of a publication. The second one is the node degree in the graph, implying that the popularity of a paper should not determine the class of a paper.

GOOD-WebKB is a university webpage network dataset. A node in the network represents a webpage, with words appearing in the webpage as node features, and edges are hyperlinks between webpages. Its 5-class prediction task is to predict the classes of webpages. GOOD Gui et al. (2022) split it according to the domain university, suggesting that classified webpages are based on word contents and link connections instead of university features.

GOOD-CBAS is a synthetic dataset modified from BA-Shapes Ying et al. (2019). The input is a graph created by attaching 80 house-like motifs to a 300-node Barabási–Albert base graph, and the task is to predict the role of nodes, including the top/middle/bottom node of a house-like motif or the node from the base graph, forming a 4-class classification task. Instead of using constant node features, GOOD Gui et al. (2022) generates colored features so that OOD algorithms need to tackle node color differences in covariate splits and color-label correlations in concept splits.

OGB-Arxiv is a dataset consists of Arxiv CS papers from 40 subject areas and their citations. The task is to predict the 40 subject areas of the papers,3 e.g., cs.AI, cs.LG, cs.OS. Instead of the semi-supervised/adaptation setting where unlabeled testing data is available during training, Wu et al. (2021) and Li et al. (2023) follow the more common and challenging out-of-distribution generalization setting, i.e., the testing nodes are not available in the training stage. Since several latent influential environment factors can change significantly over time, the properties of citation networks will vary in different time ranges. Therefore, the node distribution shifts on OGB-Arxiv are introduced by selecting papers published before 2011 as the training set, within 2011–2014 as the validation set, and within 2014–2016/2016–2018/2018–2020 as three testing sets.

F.3. Datasets in the Comparison with GCL Methods

In the seven citation networks, nodes represent papers, and edges represent citation links. In the first five networks, i.e., CoraFull, CoraML, CiteSeer, DBLP, and PubMed, given paper text as bag-of-words node features, the task is to predict the topic of a paper. In the Ogbn-Arxiv, each paper comes with a 128-dimensional feature vector obtained by averaging the embeddings of words in its title and abstract. The embeddings of individual words are computed by running the skip-gram model Mikolov et al. (2013) over the MAG corpus. The task is to predict the primary categories of the arXiv papers, which is one of the 40 subject areas of arXiv CS papers, e.g., cs.AI, cs.LG, and cs.OS, which is manually determined (i.e., labeled) by the paper's authors and arXiv moderators. The dataset is split based on the publication dates of the papers. Specifically, it is supposed to train on papers published until 2017, validate on those published in 2018, and test on those published since 2019. As for Ogbn-MAG, it is originally a heterogeneous network composed of a subset of MAG Wang et al. (2020). In this paper, we only keep the paper nodes and the citation relationships in the original graph. Similar to Ogbn-Arxiv, each paper is associated with a 128-dimensional word2vec Mikolov et al. (2013) feature vector. The task is to predict the venue (conference or journal) of each paper. In total, there are 349 different venues in Ogbn-MAG. The dataset is split using the same strategy as the Ogbn-Arxiv dataset, i.e., training models to predict venue labels of all papers published before 2018, validating and testing the models on papers published in 2018 and since 2019, respectively.

In the two product networks, nodes represent products in Amazon and edges represent that two goods are frequently bought together. Given product reviews as bag-of-words node features, the task is to map goods to their respective product category.

In the image network, nodes represent images in the Flickr website and edges represent that two images share some common properties (e.g., same geographic location and comments by the same user, etc.). Given the bag-of-word representation of the images as node features, the task is to predict the type of an image.

The graph statistics of the above datasets are shown in Table F.7 below. The last column represents the number of the investigated split types in this work; for example, Cora_Full is involved with 6 split types: IID split, Node2Vec-based clustering-based OOD split, word- or degree-based covariate, or concept OOD split.

Table F.7: Datasets statistics

DATASET	#NODES	#EDGES	#FEATURES	#CLASSES	#SPLIT TYPES
CORA_FULL	19,793	126,842	8,710	70	6
CORA_ML	2,995	16,316	2,879	7	2
CITESEER	4,230	10,674	602	6	2
DBLP	17,716	105,734	1,639	4	2
PUBMED	19,717	88,648	500	3	2
COMPUTERS	13,752	491,722	767	10	2
PHOTO	7,650	238,162	745	8	2
FLICKR	89,250	899,756	500	7	2
OGBN-ARXIV	169,343	2,315,598	128	40	4
OGBN-MAG	736,389	10,792,672	128	349	1
CBAS	700	3,962	4	4	2
WEBKB	617	1,138	1703	5	2

Moreover, the distribution of sample categories and the homophily ratio of the datasets are shown in Table F.8.

G. Baseline Details

G.1. Graph OOD Algorithms

Traditional invariant learning methods:

1. IRM Arjovsky et al. (2019): It is a representative invariant learning method. To learn invariances across environments for enabling OOD generalization, it seeks to find data representations or features so the optimal classifier on top of that representation matches all environments.
2. VREx Krueger et al. (2021): This method is proven to be able to recover the causal mechanisms of the targets and is robust to distribution shifts. Specifically, it minimizes the risk variances of the training environments for reducing the risk variances of the test environments, leading to good OOD generalization.
3. GroupDRO Sagawa et al. (2019): It introduces a new stochastic optimizer for group distributional robust optimization that is stable and scales to large models and datasets.

Domain generalization:

Table F.8: Datasets Class Distribution Statistics

Dataset	Number of Classes	% of the Most Frequent Class	% of the Least Frequent Class	Homophily Ratio
Twitch	2	55.62	44.38	60.25%
WebKB	5	34.85	11.02	15.38%
CBAS	4	42.86	11.43	79.81%
Arxiv	40	16.13	0.02	65.42%
MAG	349	4.20	0.03	30.29%
Cora	70	4.69	0.08	56.70%
Cora_ML	7	28.61	6.44	78.86%
CiteSeer	6	19.65	13.19	94.94%
DBLP	4	44.71	11.19	82.79%
PubMed	3	39.94	20.81	80.24%
Computers	10	37.51	2.12	77.72%
Photo	8	25.37	4.33	82.72%
Flickr	7	42.26	3.90	31.95%

1. DANN Ganin et al. (2016): This approach is based on the theory of domain adaptation suggesting that, for effective domain transfer to be achieved, predictions must be made based on features that cannot discriminate between the training (source) and test (target) domains.
2. DeepCoral Sun and Saenko (2016): It extends CORAL Sun et al. (2016) to learn a nonlinear transformation that aligns correlations of layer activations in deep neural networks.

Graph augmentation:

1. Mixup Wang et al. (2021): It proposes the two-branch graph convolution to mix the receptive field subgraphs for the paired nodes and enable GNNs to learn more discriminative features and reduce over-fitting.

Graph OOD Methods:

1. SRGNN Zhu et al. (2021a): It is designed to account for distributional differences between biased training data and the graph's true inference distribution. It adapts GNN models for the presence of distributional shifts between the nodes that have had labels provided for training and the rest of the dataset.
2. EERM Wu et al. (2021): It is a recent pioneering work that can tackle node-level prediction tasks under distribution shifts and achieves a valid solution for the node-level OOD problem under mild conditions. It studies invariant predictions on the graph by assuming all nodes share a single environment.
3. FLOOD Liu et al. (2023b): It is a flexible invariant Learning framework that comprises two key components, invariant learning and bootstrapped learning.
4. GIL Li et al. (2022b): It learns invariant graph-level representations under distribution shifts.
5. INL Li et al. (2023): It is a novel invariant node representation learning approach capable of generating invariant node representations based on the invariant patterns under distribution shifts with multiple latent environments by leveraging the invariance principle.

G.2. GCL Algorithms

GCL methods with uniform or adaptive data augmentation:

1. DGI (Velickovic et al., 2019): Deep Graph Infomax, a general approach reling on maximizing mutual information between patch representations and corresponding high-level summaries of graphs—both derived using established graph convolutional network architectures. The learnt patch representations summarize subgraphs centered around nodes of interest, and can thus be reused for downstream node-wise learning tasks.

2. MVGRL (Hassani and Ahmadi, 2020): Multi-View Graph Representation Learning, an approach contrasting encodings from first-order neighbors and a general graph diffusion and also contrasting node and graph encodings across views.
3. Grace (Zhu et al., 2020): GRAph Contrastive rEpresentation learning, an approach generating two graph views by corruption and learn node representation by maximizing the agreement of node representations in these two views.
4. GCA (Zhu et al., 2021b): Graph Contrastive representation learning with Adaptive augmentation, an approach designing augmentation scheme based on node centrality measures to highlight important connective structures.
5. BGRL (Thakoor et al., 2021): Bootstrapped Graph Latents, an graph representation learning method that learns by predicting alternative augmentations of the input. BGRL uses only simple augmentations and alleviates the need for contrasting with negative examples, and is thus scalable by design.
6. CCA-SSG (Zhang et al., 2021a): Canonical Correlation Analysis inspired Self-Supervised Learning on Graphs, an approach generating two views of an input graph through data augmentation and optimizing an innovative feature-level objective inspired by classical canonical correlation analysis.

GCL methods based on rational/saliency:

1. RGCL (Li et al., 2022c): Rationale-aware Graph Contrastive Learning, an unsupervised approach using a rationale generator to reveal salient structures about graph instance-discrimination as the rationale, and then creating rationale-aware views for contrastive learning. Note that this method, designed for graph property prediction tasks, integrates the views generation module and the inference flow of the predictor. Therefore, we regard node property prediction tasks as ego-graph property prediction tasks to adapt to this method.
2. CGC (Yang et al., 2023a): A novel method to utilize counterfactual mechanism to generate artificial hard negative samples for graph-level contrastive learning. It ensures that the generated negative samples are similar to the raw sample in the structure, but can have different (self-unsupervised) labels from the raw sample.
3. GCIL (Mo et al., 2024): Graph Contrastive Invariant Learning, an approach studying GCL from the perspective of causality, which introduces the spectral graph augmentation to simulate the intervention upon non-causal factors and designs the invariance objective and independence objective to better capture the causal factors.

Supervised GCL methods:

1. AutoGCL (Yin et al., 2022): Automated Graph Contrastive Learning, an approach employing a set of learnable graph view generators orchestrated by an auto augmentation strategy, where every graph view generator learns a probability distribution of graphs conditioned by the input. This method is proposed for graph property prediction tasks. However, it can be directly transferred to node property tasks since its views generator and task predictor are separated.

H. Implementation Details

General Configuration. All the experiments were run 5 times with random seeds from 0 to 4 on a Ubuntu 18.04 server with one Nvidia Tesla V100-32G GPU. And the code was implemented using Python 3.8 with PyG 2.0.4 and Pytorch 1.11 which used CUDA version 11.3. For all datasets, the maximum number of sampled neighbors per layer was set to 64. The batch size was set to 64 for all models in all datasets (except Ogbn-Arxiv and Ogbn-MAG). In Ogbn-Arxiv and Ogbn-MAG, the batch size was set to 1024. The AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 0.01 was used to train all models. The weight of the contrastive loss in the comparison of GCL methods was searched from 0.1 to 0.9. Moreover, for all baselines with hyperparameters of edge drop probabilities and temperature, we searched edge drop probabilities over [0.1, 0.2, 0.3, 0.4] and searched temperatures over [0.1, 0.2], unless the original paper reported the best choices on the datasets. As for CCL-Gn, τ in Equation 10 was searched over [0.05, 0.1, 0.15, 0.2].

Open-source Code Claim. All of the codes including dataset processing procedures, model construction, and training pipeline will be made public.

I. Additional Experimental Results

I.1. Why can't MVGRL and CGC scale to large datasets?

MVGRL's time complexity is primarily due to its use of the graph diffusion scheme for positive view augmentation. This algorithm calculates the propagation probability between nodes iteratively, resulting in a high time complexity. Additionally, the adjacency matrix generated by the graph diffusion is dense, further increasing the time and space complexity of the Graph Neural Networks (GNNs) used to encode the graph. CGC's time complexity stems from the need to initialize and optimize a specific trainable adjacency matrix for each input graph individually in order to generate the negative view. This process is time and memory-intensive. Furthermore, during the training phase, CGC still requires loading the adjacency matrix of all negative views, leading to significant memory overhead, especially for large-scale and dense graphs.

I.2. In-distribution Experimetnal Results

Table I.9: Comparison with GCL Algorithms: Node classification accuracy on independently and identically distributed test sets. All the GCL methods (include CCL-Gn) function as an auxiliary learning objective as suggested in Xie et al. (2022). GraphSAGE or GIN were selected as the base GNNs.

Method		CiteSeer	Computers	Cora_Full	Cora_ML	DBLP	Flickr	Photo	PubMed
SAGE	ERM	91.25±0.28	85.62±2.05	55.72±1.17	83.83±1.48	80.97±0.99	42.38±0.24	92.99±0.35	85.34±0.70
	MVGRL	65.83±1.48	OOM	OOM	44.36±7.12	74.19±0.97	OOM	OOM	86.16±0.71
	Grace	89.83±0.45	86.67±1.40	56.47±0.88	84.13±1.57	81.74±0.71	49.30±0.56	93.48±1.07	86.86±0.39
	GCA	90.20±0.88	84.71±1.74	56.87±1.34	84.52±0.90	81.81±0.56	49.76±1.00	92.50±1.04	86.85±0.33
	BGRL	90.17±1.00	86.25±0.65	56.40±1.25	83.64±2.43	81.35±0.32	50.69±0.25	91.90±1.83	86.50±0.29
	RGCL	89.29±0.87	58.31±12.38	49.00±1.89	80.68±1.66	80.24±0.61	38.91±7.54	81.26±7.22	81.85±0.52
	CGC	89.98±0.88	OOM	OOM	84.74±1.58	OOM	OOM	OOM	86.29±0.57
	AutoGCL	90.26±0.39	85.73±0.88	56.30±0.72	83.98±0.79	81.33±1.21	40.99±1.80	92.40±0.93	86.45±0.69
	CCL-Gn	**91.65±0.26**	**88.65±0.47**	**58.10±0.91**	**85.19±1.06**	**84.01±0.46**	**51.14±0.15**	**94.52±0.25**	**88.17±0.39**
GIN	ERM	89.36±0.80	82.97±2.20	56.40±0.61	81.84±0.96	79.52±0.74	45.31±2.56	88.25±1.66	86.05±0.41
	MVGRL	85.90±1.14	OOM	OOM	71.27±3.32	81.47±0.76	OOM	OOM	85.18±0.78
	Grace	88.74±0.85	88.35±0.86	55.96±0.68	83.01±0.95	79.64±1.00	44.11±1.19	93.21±0.69	86.83±0.21
	GCA	89.53±0.22	85.96±1.69	56.45±0.89	83.77±1.03	80.12±0.81	42.76±1.18	92.75±1.49	86.80±0.51
	BGRL	89.93±0.85	88.45±1.34	56.83±0.85	82.98±1.05	80.66±0.97	43.49±1.15	91.95±1.47	86.86±0.26
	RGCL	88.50±1.36	82.95±1.80	52.19±0.75	80.42±0.83	79.85±0.62	43.90±2.66	89.34±1.06	81.61±0.25
	CGC	89.76±1.34	OOM	OOM	82.68±2.09	OOM	OOM	OOM	84.46±1.03
	AutoGCL	89.17±1.11	88.46±1.87	55.21±0.89	82.39±1.34	79.91±0.55	37.66±1.93	92.40±1.87	86.69±0.92
	CCL-Gn	**90.98±0.56**	**90.36±0.54**	**58.46±0.70**	**84.52±0.72**	**83.42±0.23**	**51.92±0.17**	**94.57±0.06**	**87.66±0.23**

I.3. Out-of-distribution Experimetnal Results

The detailed experimental results under the out-of-distribution (OOD) setting are shown in Table I.10. It's observed that CCL-Gn still significantly enhanced the performance of base GNNs up to 18.26% with an average improvement of 6.02% and consistently outperformed the SOTA methods up to 7.40% with an average improvement of 1.62%. While a larger performance degradation suggests a larger distribution gap between the test set and the training/validation set, it's observed that base GNNs can benefit more from CCL-Gn compared to SOTA methods. For example, Cora_Full, Photo, and Flickr are the three datasets with the most significant performance degradation and CCL-Gn outperformed the best SOTA the most on these three datasets. All these results demonstrate the superiority of CCL-Gn to enhance the generalizability of various GNNs.

J. Additional Discussion on the Comparison with the GCL Baselines

Based on the experimental results of the baseline model, we can further speculate on the reasons for the success of our model.

Table I.10: Comparison with GCL Algorithms: Node classification accuracy on out-of-distribution test sets. All the GCL methods (include CCL-Gn) function as an auxiliary learning objective as suggested in Xie et al. (2022).

Method		CiteSeer	Computers	Cora_Full	Cora_ML	DBLP	Flickr	Photo	PubMed
	Δ	-1.21%	4.76%	-46.81%	-5.18%	15.80%	-12.31%	-19.19%	-0.09%
	↑	1.08%	0.88%	2.89%	0.54%	0.80%	2.95%	3.05%	0.74%
SAGE	ERM	90.29±0.68	82.89±12.99	31.88±5.88	80.05±4.98	91.22±0.83	46.66±0.37	75.57±6.13	87.26±0.52
	MVGRL	68.41±2.60	OOM	OOM	51.24±6.09	86.90±1.43	OOM	OOM	87.65±0.73
	Grace	89.77±0.48	91.15±1.12	32.34±2.88	83.24±3.70	91.67±0.52	46.58±0.56	77.65±5.59	87.88±0.56
	GGA	89.77±1.60	88.85±3.26	33.78±4.04	83.64±2.92	91.97±1.46	46.29±0.43	76.12±8.19	88.06±0.62
	BGRL	89.65±0.73	88.58±3.45	33.93±2.19	83.11±3.58	92.42±0.58	46.55±0.30	80.22±6.55	88.11±0.87
	RGCL	89.29±1.69	51.95±40.35	25.46±4.34	79.63±5.02	91.99±0.99	36.80±0.10	49.18±28.98	86.86±0.33
	CGC	89.62±0.30	OOM	OOM	83.18±2.72	OOM	OOM	OOM	87.89±0.94
	AutoGCL	89.61±0.75	86.44±7.83	32.12±1.32	82.27±5.09	91.74±0.60	36.69±4.84	74.76±3.27	87.02±0.58
	CCL-Gn	**91.15±1.03**	**92.62±0.23**	**36.44±3.38**	**84.36±3.99**	**93.31±0.30**	**47.51±0.24**	**84.42±2.11**	**88.77±0.39**
GAT	ERM	89.98±0.91	89.98±4.48	36.99±3.82	83.30±2.96	91.70±1.13	43.96±4.17	74.74±9.61	87.46±1.06
	MVGRL	88.45±0.96	OOM	OOM	81.96±3.96	92.07±1.04	OOM	OOM	87.81±0.63
	Grace	88.93±1.00	93.15±1.08	39.87±4.66	84.19±2.74	93.14±0.32	**47.38±0.60**	86.14±2.50	87.74±0.34
	GGA	90.05±0.97	93.27±0.49	**40.28±1.63**	84.75±2.73	92.96±0.45	46.79±0.44	**86.80±3.12**	88.26±0.29
	BGRL	90.37±0.60	91.96±0.79	39.49±3.42	83.81±3.45	92.74±0.68	44.94±0.68	81.15±3.33	88.04±0.69
	RGCL	88.37±1.61	55.5±37.86	22.75±3.24	79.01±4.49	92.40±0.71	37.67±3.49	66.25±10.34	86.94±0.36
	CGC	90.25±1.03	OOM	OOM	83.83±2.49	OOM	OOM	OOM	87.61±0.39
	AutoGCL	89.96±0.64	91.97±1.02	37.96±3.01	82.67±4.03	92.41±0.42	46.20±2.46	85.35±3.06	87.44±0.48
	CCL-Gn	**91.44±0.75**	**93.29±0.45**	40.19±2.00	**85.35±1.84**	**93.40±0.50**	47.22±0.35	85.98±1.73	**88.62±0.32**
GIN	ERM	88.97±1.17	90.15±1.84	31.24±3.57	80.51±5.39	89.93±1.16	44.54±2.47	72.64±10.94	88.17±0.61
	MVGRL	86.85±1.82	OOM	OOM	69.85±6.36	92.02±1.24	OOM	OOM	87.64±0.75
	Grace	89.57±1.02	92.81±0.87	32.82±1.03	82.98±3.37	86.43±1.42	38.11±7.02	81.99±2.22	87.72±0.77
	GGA	89.29±1.33	90.47±2.27	33.59±1.02	83.11±3.53	88.62±1.34	41.04±1.11	81.21±4.00	87.98±0.60
	BGRL	89.49±0.49	91.96±0.73	34.74±2.35	82.29±3.29	91.02±1.37	38.22±5.15	77.82±10.65	88.19±0.54
	RGCL	88.08±1.00	86.98±2.55	31.11±2.97	81.96±2.34	92.18±0.53	41.46±3.70	80.43±1.63	86.72±0.45
	CGC	89.74±1.10	OOM	OOM	79.58±4.85	OOM	OOM	OOM	87.22±0.83
	AutoGCL	87.68±1.30	90.48±0.93	28.27±5.45	78.89±4.57	88.54±0.64	33.57±1.67	77.12±13.32	87.19±2.11
	CCL-Gn	**90.56±0.45**	**93.75±0.53**	**35.26±1.00**	**83.16±3.13**	**93.26±0.35**	**47.82±0.41**	**85.98±1.03**	**89.13±0.34**

[1] Δ: the average accuracy degradation of base GNNs on the testing set compared to the accuracy on the validation set.

[2] ↑: the average gain of CCL-Gn against the best SOTA GCL methods.

1. **The importance of downstream task dependencies.** It's observed that unsupervised augmentation strategies sometimes even reduce the task performance of the models in the independently and identically distributed (iid) setting. Even CGC, which also absorbs the theory of counterfactual, is still unable to effectively improve the performance of the model. This might support our discussion of downstream task dependency in the introduction. Unexpectedly destroying the integrity of the task-relevant structures can make the model learn task-oriented node representations in the wrong direction.
2. **The importance of sufficient decomposition of causal and spurious features.** AutoGCL met a performance degradation from the iid setting to the out-of-distribution (ood) setting. It achieved competitive performance in the iid setting and sometimes could be the best baseline. Yet, it displayed a very inferior performance in the ood setting. This can verify the discussion in our introduction, i.e., simply discovering the predictive subgraph can strengthen the spurious structures thus compromising the generalizability of the models.
3. **The capacity of CCL-Gn to auto-discover important structures.** GCA, an approach that adopts adaptive data augmentation and thus enables models to learn important structures from the perspective of network science, outperforms other baselines in most settings. This can imply the effectiveness of CCL-Gn to automatically sort the vital structures in the positive view without human design.
4. **The necessity of ad-hoc design for node property prediction tasks.** We found that RGCL failed in most settings and sometimes can get a great performance variance. As we know, RGCL is designed for graph property prediction tasks and integrates the views generation module and the inference flow of the predictor. Therefore, RGCL can only do the node property prediction tasks as ego-graph property prediction tasks. From

Table K.11: Real runtime (seconds per epoch) comparison. **Bold** and underline indicate the fastest the top-3 fastest speed, respectively.

Dataset	Base	MVGRL	Grace	GCA	BGRL	RGCL	AutoGCL	CCL-Gn
CiteSeer	2.04±0.09	2.50±0.41	1.75±0.10	3.05±0.16	2.80±0.57	**1.71±0.07**	2.03±0.13	2.61±0.52
Computers	5.02±0.11	OOM	13.13±2.64	10.35±0.19	12.16±0.23	53.40±12.15	**8.16±0.49**	9.95±0.22
Cora_Full	8.41±0.11	OOM	29.43±6.99	22.80±4.71	31.25±11.73	31.60±0.47	12.60±1.17	**12.31±0.94**
Cora_ML	2.10±0.17	3.25±0.12	2.60±0.46	2.87±0.74	2.71±0.16	2.12±0.06	**2.05±0.07**	2.36±0.54
DBLP	3.66±0.17	23.36±0.53	9.67±0.70	16.17±5.19	9.22±0.55	11.15±0.57	5.04±0.15	**4.92±0.85**
Flickr	9.81±0.59	TLE	28.35±3.13	46.36±15.96	59.72±36.26	110.66±15.47	17.91±1.00	**14.93±2.43**
Photo	3.06±0.06	OOM	5.64±1.28	4.45±0.07	5.53±0.09	15.35±2.49	**4.01±0.11**	4.42±0.18
PubMed	3.13±0.15	27.12±1.32	8.20±1.23	7.37±1.59	8.10±1.22	10.02±0.19	**4.49±0.26**	4.56±0.42

[1] OOM means Out Of Memory (>32GB) and TLE means Time Limit Exceeded (seconds per epoch > 1000s)

this view, we might conclude that the failure of RGCL is due to the task shift from graph classification to node classification. This can highlight the importance of ad-hoc design for node tasks, at least enabling the decoupling of the contrastive view generation and the node prediction modules. Similarly, AutoGCL, while being the only task-oriented baseline, can not achieve better performance against some unsupervised augmentations even in partial settings in the independently and identically distributed test sets. This might also indicate that the node task and the graph task are substantially different and can benefit from ad-hoc design.

K. Real runtime evaluation

The real runtime of a model on a dataset was calculated by averaging the minimum one-epoch runtime of the three base GNNs at all settings on the dataset, which is displayed in Table K.11. Note that all the experiments were conducted in the same environment described in Section **??**. **It's seen that CCL-Gn actually achieves a very fast speed compared to other baselines.** Especially when the graph becomes larger and more complex, e.g. on the Flickr dataset, CCL-Gncan achieve the fastest speed. This verified the statement in complexity analysis and scalability discussion, which suggest the following two reasons: (a) CCL-Gnleverages a simple edge probability estimator compared with other learnable data augmentation methods in previous GCL works; (b) There is no need to contrast in-batch samples since each sample is associated with a negative counterfactual view. The real runtime of a model on a dataset is calculated by averaging the minimum one-epoch runtime of the three base GNNs at all settings on the dataset (As shown in Table K.11). **It's seen that CCL-Gn achieves a very fast speed compared to other baselines.** Especially when the graph becomes larger and more complex, e.g. on the Flickr dataset, CCL-Gncan achieve the fastest speed. This verified the statement in complexity analysis and scalability discussion, which suggest the following two reasons: (a) CCL-Gnleverages a simple edge probability estimator compared with other learnable data augmentation methods in previous GCL works; (b) There is no need to contrast in-batch samples since each sample is associated with a negative counterfactual view.