Disentangled Multiplex Graph Representation Learning

Yujie Mo [1] **Yajie Lei** [1] **Jialie Shen** [2] **Xiaoshuang Shi** [1] **Heng Tao Shen** [1] **Xiaofeng Zhu** [1 3]

Abstract

Unsupervised multiplex graph representation learning (UMGRL) has received increasing interest, but few works simultaneously focused on the common and private information extraction. In this paper, we argue that it is essential for conducting effective and robust UMGRL to extract complete and clean common information, as well as more-complementarity and less-noise private information. To achieve this, we first investigate disentangled representation learning for the multiplex graph to capture complete and clean common information, as well as design a contrastive constraint to preserve the complementarity and remove the noise in the private information. Moreover, we theoretically analyze that the common and private representations learned by our method are provably disentangled and contain more task-relevant and less task-irrelevant information to benefit downstream tasks. Extensive experiments verify the superiority of the proposed method in terms of different downstream tasks.

1. Introduction

Recently, multiplex graph representation learning (MGRL) has emerged as a powerful tool to model complex relationships among nodes (Chu et al., 2019; Zhang & Kou, 2022). In particular, unsupervised multiplex graph representation learning (UMGRL) methods have attracted increasing attention due to the label availability for the training process and have shown great potential in a wide range of applications, such as anomaly detection and recommendation systems (Chen et al., 2022; Xie et al., 2022).

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China [2]Department of Computer Science, City, University of London, London, United Kingdom [3]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China. Correspondence to: Xiaofeng Zhu <seanzhuxf@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Existing UMGRL methods can be roughly divided into two categories, *i.e.,* traditional unsupervised methods and self-supervised methods. Traditional UMGRL methods aim to extract hidden information without the help of labels by random walk strategies (Dong et al., 2017) or proximity preservation (Shi et al., 2018a). However, they often overemphasize proximity information and ignore node features. Inspired by the prosperity of self-supervised learning (He et al., 2020; Chen & He, 2021), self-supervised MGRL methods (including intra-graph and inter-graph contrastive learning methods) have been developed and achieved great success in recent years. The intra-graph contrastive learning methods aim to capture the global properties, but they ignore intrinsic associations among different graphs (Park et al., 2020) and may lead to suboptimal representations. To solve this issue, inter-graph contrastive learning methods are proposed as an alternative by modelling common information among different graphs (Zhou et al., 2022). The common information contains the consistency among all graphs and has been verified to be the key component of sample identification (Zhu et al., 2022).

Although existing UMGRL especially inter-graph contrastive learning methods have achieved promising performance in many tasks, there are still some limitations to be addressed. On the one hand, previous UMGRL methods are designed to implicitly capture the common information, but such a process is performed in a black box (Zhu et al., 2022; Li et al., 2022). As a result, previous UMGRL methods cannot verify if all common information has been obtained (*i.e.,* complete) and if it contained other confusing contents (*i.e.,* clean). On the other hand, apart from the common information, the rest of the content specific to each graph can be referred to as private information. Some private information is noise, but the other parts contain complementary information, which is different from that in other graphs and has been demonstrated to facilitate downstream tasks in many fields (Xie et al., 2020; Wang et al., 2022a). However, previous UMGRL methods do not consider private information and thus ignoring its complementarity (Zhu et al., 2022). Moreover, they generally fuse all representations from different graphs and thus including noise into the fusion process (Jing et al., 2021a).

Based on the above observations, it is a possible solution to improve the effectiveness and robustness of UMGRL by ex-

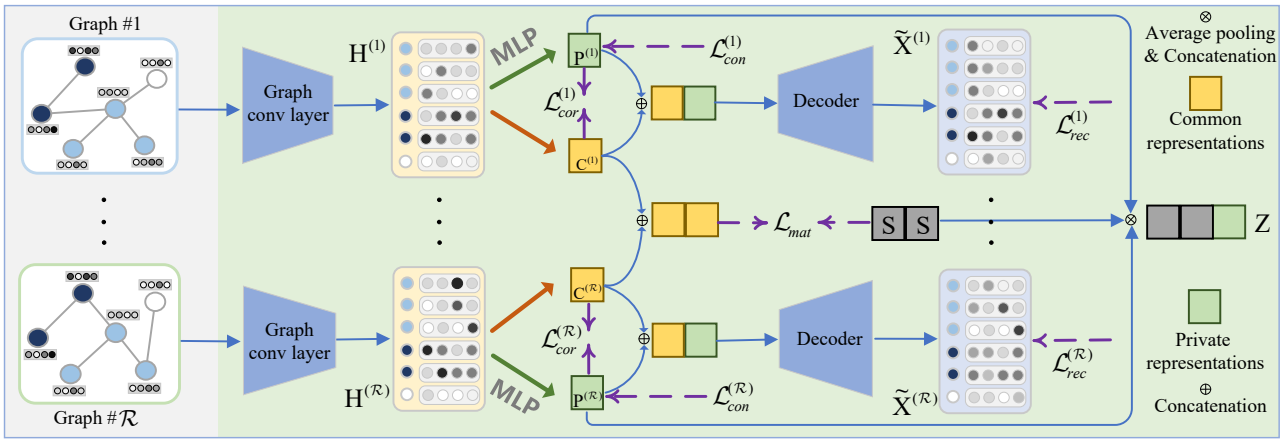


Figure 1. The flowchart of the proposed DMG. Specifically, given the node feature matrix and graph structures, DMG first employs the graph convolutional layer and Multi-Layer Perceptron (MLP) to generate common and private representations (*i.e.,* $\mathbf{C}^{(r)}$ and $\mathbf{P}^{(r)}$) for every graph. After that, DMG investigates the matching loss $\mathcal{L}_{mat}$ and the correlation loss $\mathcal{L}_{cor}^{(r)}$, respectively, to obtain the complete and the clean common information, as well as investigate the reconstruction loss $\mathcal{L}_{rec}^{(r)}$ to promote the invertibility of encoders for exploring the issue of the trivial solution. Meanwhile, DMG investigates the contrastive loss $\mathcal{L}_{con}^{(r)}$ to preserve the complementarity and remove the noise in the private information. Finally, private representations of different graphs are first fused by the average pooling and then concatenated with the common variable $\mathbf{S}$ to obtain the final representations $\mathbf{Z}$.

plicitly capturing complete and clean common information, as well as preserving complementarity and removing noise in the private information. To achieve this, there are two crucial challenges to be solved, *i.e.,* (i) it is difficult to decouple the common information from private information as they are generally mixed together, and (ii) it is necessary to distinguish and further preserve the complementarity from noise in the private information.

In this paper, to address the above issues, different from previous traditional UMGRL and self-supervised MGRL, we investigate a new unsupervised framework, *i.e.,* **D**isentangled **M**ultiplex **G**raph representation learning (DMG for brevity), to conduct effective and robust UMGRL, as shown in Figure 1. To do this, we first decouple the common and private representations by designing a new disentangled representation learning for the multiplex graph to extract complete and clean common information, and thus tackling Challenge (i). Moreover, we further preserve the complementarity and remove the noise in the private information by designing a contrastive constraint on private representations, to tackle Challenge (ii). As a result, the common and private representations learned by our method can be provably disentangled and contain more task-relevant and less task-irrelevant information to benefit downstream tasks.

Compared to previous UMGRL methods, the main contributions of our method can be summarized as follows:

- To the best of our knowledge, we make the first attempt to disentangle the common and private representations for the multiplex graph to extract complete and clean common information. We further propose a contrastive constraint to preserve the complementarity and remove the noise in the private information.
- We theoretically prove that the representations learned by our method can extract complete and clean common information. We further prove that the common and private representations learned by our method contain more task-relevant information and less task-irrelevant information.
- We experimentally demonstrate the effectiveness and robustness of the proposed method in terms of node classification and node clustering tasks on multiplex graph datasets and single-view graph datasets, compared to numerous comparison methods.

2. Related Work

This section briefly reviews the topics related to this work, including unsupervised multiplex graph representation learning and disentangled representation learning.

2.1. Unsupervised Multiplex Graph Representation Learning

Unsupervised multiplex graph representation learning (UMGRL) has emerged as a popular method and has drawn considerable attention in recent years as it eliminates the reliance on label information, and the capability to model the complex relationships between nodes (Chu et al., 2019). Existing UMGRL methods can be classified into two cat-

egories, *i.e.,* traditional unsupervised methods and self-supervised methods. Traditional unsupervised methods tend to generate representations by random walk strategies (Dong et al., 2017) or proximity preservation (Shi et al., 2018a) based on the proximity among nodes. For example, MNE (Zhang et al., 2018) performs meta-path based random walks to extract information of multi-type relations into a unified representations space. Similarly, HERec (Shi et al., 2019) first designs a type constraint strategy to filter the node sequence and then learns representations by random walk based strategies. As for proximity preservation, AspEm (Shi et al., 2018a) alleviates the incompatibility among different graphs by preserving a set of representative proximity information in the multiplex graph. HEER (Shi et al., 2018b) further improves the effectiveness by coupling the edge representations with inferred type-specific metrics.

Limited by the overemphasis on proximity information and node feature ignorance of traditional unsupervised methods, self-supervised MGRL methods have shown remarkable performance. Existing self-supervised MGRL methods can be divided into two subgroups, *i.e.,* intra-graph contrastive learning methods and inter-graph contrastive learning methods. For example, DMGI (Park et al., 2020) and HDMI (Jing et al., 2021a) conduct contrastive learning by maximizing the mutual information between node representations and the graph summary within each graph, but overlook the intrinsic correlation among different graphs. Different to the intra-graph contrastive learning, STENCIL (Zhu et al., 2022) conducts inter-graph contrastive learning by contrasting node representations from each graph and an aggregation graph. CKD (Zhou et al., 2022) adopts contrastive learning between node representations and high-level representations of different graphs under the collaborative knowledge distillation framework. Despite their success, existing methods fail to obtain complete and clean common information, as well as more-complementarity and less-noise private information, which is significant for downstream tasks.

2.2. Disentangled Representation Learning

Disentangled representation learning aims to learn representations that identify and disentangle the underlying explanatory factors hidden in the given data. Existing disentangled representation learning has already made promising developments with great potential based on the generative models (Higgins et al., 2016; Xu et al., 2021a; Xiao et al., 2022). For example, Beta-VAE (Higgins et al., 2016) introduces an adjustable parameter to balance latent channel capacity and independence constraints with reconstruction accuracy and thus improving the variational auto-encoder (VAE) framework. DDPAE (Jiang et al., 2020) proposes a decompositional disentangled predictive auto-encoder framework to learn both the latent decomposition and disentanglement without explicit supervision. PSGAN (Jiang et al., 2020) proposes to disentangle the content information and style information of images to generate the style transferred images based on the generative adversarial network. S3VAE (Zhu et al., 2020a) proposes a self-supervised sequential VAE model which disentangles the time-varying variables and time-invariant variables of video and audio sequences.

Inspired by the prosperity in other fields, disentangled representation learning has recently raised a surge of interest in graph-structured data (Ma et al., 2019; Yang et al., 2020; Mercatali et al., 2022). For example, GraphVAE (Simonovsky & Komodakis, 2018) transfers the generative models for images and text to the domain of graphs and is available to output a probabilistic fully-connected graph directly based on VAE framework. GraphLoG (Xu et al., 2021b) proposes to conduct self-supervised graph-level representation learning by disentangling the local similarities and global semantic clusters. DGCL (Li et al., 2021) proposes to learn disentangled graph-level representations with self-supervision by forcing the factorized representations to independently reflect the expressive information from different latent factors. DSSL (Xiao et al., 2022) decouples different underlying semantics between different neighborhoods into the self-supervised learning process based on the generative model. Although the above methods achieve excellent results on different tasks, they are all designed for the single-view graph and thus fail to take into account the complex relationships between nodes. Moreover, these methods do not consider the complementarity and noise within each graph structure leading to suboptimal performance.

3. Method

Notations. Let $\mathcal{G} = \{\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \ldots, \mathcal{G}^{(\mathcal{R})}\}$ to denote the multiplex graph, where $\mathcal{G}^{(r)} = \{\mathcal{V}, \mathcal{E}^{(r)}\}$ is the r-th graph in the multiplex graph, $\mathcal{R}$ is the number of graphs. $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ and $\mathcal{E}^{(r)}$ represent the node set of all graphs and the edge set of the r-th graph, respectively. We denote node features of each graph as $\mathbf{X}^{(r)} = \mathcal{T}(\mathbf{X}) \in \mathbb{R}^{N \times F}$, where $\mathcal{T}$ denotes the random dropout operation, $\mathbf{X}$ denotes original node features, N and F denote the number of nodes and the dimension of node features, respectively. $\mathbf{A}^{(r)} \in \mathbb{R}^{N \times N}$ denotes the graph structure of the r-th graph, where $\mathbf{A}^{(r)}_{ij} = 1$ iff $e^{(r)}_{ij} = (v_i, v_j) \in \mathcal{E}^{(r)}$. The proposed DMG first learns the disentangled common representations $\mathbf{C}^{(r)} \in \mathbb{R}^{N \times D}$ and private representations $\mathbf{P}^{(r)} \in \mathbb{R}^{N \times d}$ through a common variable $\mathbf{S} \in \mathbb{R}^{N \times D}$, and then obtain the fused representations $\mathbf{Z} \in \mathbb{R}^{N \times (\mathcal{R} \times D + d)}$, where D and d are the dimensions of the representation space.

3.1. Motivation

Previous UMGRL methods aim to implicitly extract common information among different graphs, which is effective and robust in revealing the identity of samples (Zhou et al.,

2022; Zhu et al., 2022). However, they generally ignore the complementarity in the private information of each graph, and may lose significant properties among nodes. For example, in a multiplex graph, where papers are nodes and edges represent either co-subjects or co-authors in two different graphs. If a private edge (*e.g.,* the co-subject relation) exists only within a certain graph and interconnects two papers from the same class, it benefits to reduce the intra-class gap by providing complementary information to identify papers. Therefore, it is necessary to consider both the common information and the private information for achieving an effective and robust UMGRL.

Based on the common information that helps to identify the samples, it is intuitive to capture all the common information among different graphs (*i.e.,* complete). Moreover, such complete common information should contain common information only (*i.e.,* clean). In contrast, if common information contains other confusing contents, the quality of the common information can be compromised. Therefore, the first question comes: *How to obtain complete and clean common information?* On the other hand, private information is a mix of complementarity and noise. Considering the same example of citation networks, if a private edge interconnects two papers from different classes, it can interfere the message passing and should be removed as noise. Therefore, the second question comes: *How to preserve complementarity and remove noise in private information?*

However, few previous UMGRL methods explored the above questions. Recently, disentangled representation learning methods have been developed to obtain common and private representations (Ma et al., 2019; Li et al., 2021; Lyu et al., 2022; Wang et al., 2022c;b; Xiao et al., 2022), but it is challenging to apply them to solve the above issues in UMGRL due to the complex relationships among nodes in the multiplex graph, as well as the complementarity and noise in the graph structure. To do this, we propose a new disentangled multiplex graph representation learning framework, to answer the above two questions, *i.e.,* Section 3.2 for the first question, and Section 3.3 for the second question.

3.2. Common Information Extraction

Previous UMGRL methods (*e.g.,* inter-graph contrastive learning methods) generally implicitly capture the common patterns among different graphs by maximizing the mutual information between two graphs. For instance, to extract the common information, STENCIL (Zhu et al., 2022) maximizes the mutual information between each graph and the aggregation graph, while CKD (Zhou et al., 2022) maximizes the mutual information between regional representations and global representations in different graphs. However, these efforts cannot capture complete and clean common information provably as they fail to decouple the common information from private information. To solve this issue, in this paper, we investigate disentangled representation learning to obtain complete and clean common information.

Specifically, we first employ the graph convolutional layer $g^{(r)}$ to generate node representations $\mathbf{H}^{(r)}$ based on the node features and the graph structure of each graph, *i.e.,*

$$\mathbf{H}^{(r)} = \sigma(\hat{\mathbf{D}}_r^{-\frac{1}{2}}\hat{\mathbf{A}}^{(r)}\hat{\mathbf{D}}_r^{-\frac{1}{2}}\mathbf{X}^{(r)}\Theta^{(r)}), \tag{1}$$

where $\hat{\mathbf{A}}^{(r)} = \mathbf{A}^{(r)} + w\mathbf{I}_N$, and w indicates the weight of self-connections. $\hat{\mathbf{D}}_r$ is the degree matrix of $\hat{\mathbf{A}}^{(r)}$, σ is the activation function, and $\Theta^{(r)}$ is the weight matrix of $g^{(r)}$. To facilitate the decoupling of the common and private information within each graph, we then employ MLP with unshared parameters (*i.e.,* $f_c^{(r)}$ and $f_p^{(r)}$) to map node representations $\mathbf{H}^{(r)}$ of each graph into common and private representations (*i.e.,* $\mathbf{C}^{(r)}$ and $\mathbf{P}^{(r)}$).

Given $\mathbf{C}^{(1)}, ..., \mathbf{C}^{(\mathcal{R})}$, the simplest way for aligning common representations from different graphs is to directly set $\mathbf{C}^{(1)} = ... = \mathbf{C}^{(\mathcal{R})}$. However, this may affect the quality of common representations by directly aligning suboptimal common representations in the initial training process (Benton et al., 2019; Lyu & Fu, 2020; Lyu et al., 2022). In this paper, we introduce a common variable $\mathbf{S}$ with the orthogonality and zero mean via the singular value decomposition operation on common representations. After that, we conduct the matching loss between the common representations $\mathbf{C}^{(r)}$ and the common variable $\mathbf{S}$, aiming to gradually align common representations from different graphs for capturing complete common information among them. The matching loss is formulated as:

$$\begin{aligned} &\mathcal{L}_{mat} = \frac{1}{N}\sum_{r=1}^{\mathcal{R}}\sum_{n=1}^{N}(\mathbf{c}_n^{(r)} - \mathbf{s}_n)^2, \\ &\text{s.t. } \frac{1}{N}\sum_{n=1}^{N}\mathbf{s}_n\mathbf{s}_n^\top = \mathbf{I}, \frac{1}{N}\sum_{n=1}^{N}\mathbf{s}_n = \mathbf{0}. \end{aligned} \tag{2}$$

Intuitively, $\mathbf{S}$ in Eq. (2) communicates the common representations from different graphs and converges them to the consistency, *i.e.,* $\mathbf{C}^{(1)} = \ldots = \mathbf{S} = \ldots = \mathbf{C}^{(\mathcal{R})}$. Therefore, the consistency among common representations of all graphs guarantees that the common information among different graphs can be obtained completely.

After that, to decouple the common and private representations, we have to enforce the statistical independence between them. It is noteworthy that if common and private representations are statistically independent, then we have $\mathbb{E}[\phi^{(r)}(\mathbf{C}^{(r)})\psi^{(r)}(\mathbf{P}^{(r)})] = \mathbb{E}[\phi^{(r)}(\mathbf{C}^{(r)})]\mathbb{E}[\psi^{(r)}(\mathbf{P}^{(r)})]$, and vice versa, where $\phi^{(r)}$ and $\psi^{(r)}$ are measurable functions (Gretton et al., 2005). Obviously, the minimization of the correlation between $\phi^{(r)}(\mathbf{C}^{(r)})$ and $\psi^{(r)}(\mathbf{P}^{(r)})$ could

be used to achieve the independence between common and private representations. In particular, the correlation loss is obtained by calculating the Pearson's correlation coefficient between $\phi^{(r)}(\mathbf{C}^{(r)})$ and $\psi^{(r)}(\mathbf{P}^{(r)})$, *i.e.,*

$$\mathcal{L}_{cor} = \sum_{r=1}^{\mathcal{R}} \frac{|\text{Cov}(\phi^{(r)}(\mathbf{C}^{(r)}), \psi^{(r)}(\mathbf{P}^{(r)}))|}{\sqrt{\text{Var}(\phi^{(r)}(\mathbf{C}^{(r)})}\sqrt{\text{Var}(\psi^{(r)}(\mathbf{P}^{(r)})}}, \quad (3)$$

where $\text{Cov}(\cdot,\cdot)$ and $\text{Var}(\cdot)$ indicate covariance and variance operations, respectively. In Eq. (3), the correlation coefficient between $\phi^{(r)}(\mathbf{C}^{(r)})$ and $\psi^{(r)}(\mathbf{P}^{(r)})$ is encouraged to converge to 0. Actually, as the correlation loss converges, the common representations $\mathbf{C}^{(r)}$ and the private representations $\mathbf{P}^{(r)}$) are statistically independent.

Based on Eq. (3), the common representations $\mathbf{C}^{(r)}$ are expected to obtain clean common information. Therefore, we almost answer the first question by the matching loss (*i.e.,* achieving complete common information) and the correlation loss (*i.e.,* achieving clean common information). However, the learned common and private representations may be trivial solutions under the unsupervised framework (Jing et al., 2021b; Xu et al., 2022a;b). Popular solutions include contrastive learning methods and auto-encoder methods. Contrastive learning methods introduce a large number of negative samples to avoid trivial solutions, but they may induce large memory overheads (Zhang et al., 2021; Liu et al., 2023b;a). Auto-encoder methods employ the auto-encoder framework with the reconstruction loss to promote the invertibility of encoders for preventing the trivial solutions (Kipf & Welling, 2016; Liu et al., 2022). However, existing graph auto-encoders are designed to reconstruct the direct edges and ignore the topological structure as well as be with expensive computation cost (Donnat et al., 2018; Mrabah et al., 2022). To address the above issues, we investigate a new reconstruction loss to simultaneously reconstruct the node features and the topological structure.

Specifically, we first concatenate the common and private representations and then obtain the reconstructed node representations $\widetilde{\mathbf{X}}^{(r)}$ with the reconstruction network $\boldsymbol{p}^{(r)}$. We further conduct the feature reconstruction and topology reconstruction loss to reconstruct the node features and local topological structure, respectively. As a result, the reconstruction loss is formulated as:

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{r=1}^{\mathcal{R}}\sum_{n=1}^{N}((\widetilde{\mathbf{x}}_n^{(r)} - \mathbf{x}_n^{(r)})^2 + (\widetilde{\mathbf{x}}_n^{(r)} - \mathbf{x}_{n,nei}^{(r)})^2), \quad (4)$$

where $\mathbf{x}_{n,nei}^{(r)} = \frac{1}{m}\sum_{j=1}^{m}\{\mathbf{x}_j^{(r)} | v_j \in \mathcal{N}_i^{(r)}\}$, m is the number of sampled neighbors, and $\mathcal{N}_i^{(r)}$ represents the 1-hop neighborhood set of node v_i. In Eq. (4), the first term encourages $\widetilde{\mathbf{X}}^{(r)}$ to reconstruct the original node features, and the second term encourages $\widetilde{\mathbf{X}}^{(r)}$ to reconstruct the topological structure. As a result, Eq. (4) enforces that the reconstructed representations and the original input (*i.e.,* the node features and the graph topology) can be recovered from each other, thus promoting the invertibility of encoders to avoid trivial solutions.

Denoting the optimal common representations as $\mathbf{C}_*$, which contains complete and clean common information, we have Theorem 3.1 on the common information extraction. Proofs of all Theorems are shown in Appendix B.

Theorem 3.1. *Assume the solution that satisfies the constraints in Eq. (2), Eq. (3), and Eq. (4) has been found, then we have* $\mathbf{C}^{(r)} = f_c^{(r)} \circ g^{(r)}(\mathbf{X}^{(r)}, \mathbf{A}^{(r)}) = \varphi(\mathbf{C}_*)$ *for* $\forall r \in [1, \mathcal{R}]$, *where* φ *is an invertible function.*

Theorem 3.1 indicates that if the solution satisfies the constraints in Eq. (2), Eq. (3), and Eq. (4), the common representations learned by our method and the optimal common representations can be transformed from each other due to the invertibility of the function φ. Therefore, the common representations learned by our method (*i.e.,* $\mathbf{C}^{(r)}$) have all the information of the optimal common representations (*i.e.,* $\mathbf{C}_*$) and thus extracting complete and clean common information provably. As a result, based on Eq. (2), Eq. (3), and Eq. (4), we disentangle the common and private representations to obtain complete and clean common information and thus answering the first question in Section 3.1.

3.3. Private Information Constraint

Based on Section 3.1, the private information is a mix of complementarity and noise. Therefore, given the learned private representations, we hope to further answer the second question in Section 3.1, *i.e.,* to preserve the complementarity and remove the noise in the private information. Moreover, the private information of the multiplex graph mainly lies in the graph structure of each graph since node features of different graphs are generated from the shared feature matrix $\mathbf{X}$. Therefore, we investigate preserving the complementary edges and removing the noisy edges in each graph structure. To do this, we first give the following definition for complementarity and noise in graph structures:

Definition 3.2. For any private edge in the r-th graph $\mathcal{G}^{(r)}$, *i.e.,* $e_{ij}^{(r)} \in \mathcal{E}^{(r)}$, and $e_{ij}^{(r)} \notin \bigcup_{r' \in [1,\mathcal{R}], r' \neq r} \mathcal{E}^{(r')}$, if the node pair (v_i, v_j) belongs to the same class, then $e_{ij}^{(r)}$ is a complementary edge in the graph $\mathcal{G}^{(r)}$. Otherwise, $e_{ij}^{(r)}$ is a noisy edge in the graph $\mathcal{G}^{(r)}$.

Definition 3.2 divides the private information in each graph into two parts, *i.e.,* complementary edges and noisy edges, according to the classes of node pairs. However, the node labels are unavailable in an unsupervised manner. To solve this issue, in this work, we approximate the label information of the node pair (v_i, v_j) as the cosine similarity $\epsilon_{ij}^{(r)}$

between the common variables $\mathbf{s}_i$ and $\mathbf{s}_j$, *i.e.,*

$$\epsilon_{ij}^{(r)} = \frac{\mathbf{s}_i \cdot \mathbf{s}_j}{\|\mathbf{s}_i\|\|\mathbf{s}_j\|}. \tag{5}$$

Given the cosine similarity of all node pairs in the edge set $\mathcal{E}^{(r)}$, we further assume that node pairs with top similarity belong to the same class and node pairs with low similarity belong to different classes. As a result, the edges with high similarity for connected nodes are complementary edges, denoted as $\mathcal{E}_c^{(r)}$, while the edges with low similarity for connected nodes are noisy edges, denoted as $\mathcal{E}_n^{(r)}$. Intuitively, the complementary edges should be preserved while the noisy edges should be removed.

To achieve the above intuition, we design a simple contrastive module and conduct the contrastive loss between private representations of node pairs in $\mathcal{E}_c^{(r)}$ and $\mathcal{E}_n^{(r)}$, *i.e.,*

$$\mathcal{L}_{con} = -\sum_{r=1}^{\mathcal{R}} \log \frac{\sum e^{\theta\left(\mathbf{p}_i^{(r)}, \mathbf{p}_j^{(r)}\right)/\tau}}{\sum e^{\theta\left(\mathbf{p}_i^{(r)}, \mathbf{p}_j^{(r)}\right)/\tau} + \sum e^{\theta\left(\mathbf{p}_k^{(r)}, \mathbf{p}_l^{(r)}\right)/\tau}}. \tag{6}$$

$\mathbf{p}_i^{(r)}$, $\mathbf{p}_j^{(r)}$, $\mathbf{p}_k^{(r)}$, and $\mathbf{p}_l^{(r)}$ indicate the private representations of node v_i, v_j, v_k, and v_l, respectively, where $(v_i, v_j) = e_{ij}^{(r)} \in \mathcal{E}_c^{(r)}$ and $(v_k, v_l) = e_{kl}^{(r)} \in \mathcal{E}_n^{(r)}$. Moreover, θ is the cosine similarity operation and τ is a temperature parameter. Eq. (6) increases the cosine similarity of nodes connected by the complementary edges, meanwhile, it decreases the cosine similarity of nodes connected by the noisy edges. Thus Eq. (6) preserves the complementarity and removes noise in private information to answer the second question in Section 3.1. Different from the widely-used contrastive objective function, *i.e.,* the InfoNCE loss (Oord et al., 2018) regarding two augmented views of the same samples as positive pairs while our proposed contrastive loss treats two nodes connected by complementary edges as positive pairs.

Besides preserving the complementarity and removing noise, the private information constrained by Eq. (6) benefits the downstream tasks as well. Denote $\widehat{\mathbf{H}}^{(r)}$ as representations concatenated by the common and private representations learned by our method, and denote $\widetilde{\mathbf{H}}^{(r)}$ as the node representations learned by previous inter-graph contrastive learning methods, which maximize the mutual information among different graphs, we have:

Theorem 3.3. *For any downstream task T, the node representations $\widehat{\mathbf{H}}^{(r)}$ contain more task-relevant information and less task-irrelevant information than $\widetilde{\mathbf{H}}^{(r)}$, i.e.,*

$$\begin{aligned} I(\widehat{\mathbf{H}}^{(r)}, T) &\geq I(\widetilde{\mathbf{H}}^{(r)}, T), \\ H(\widehat{\mathbf{H}}^{(r)}|T) &\leq H(\widetilde{\mathbf{H}}^{(r)}|T), \end{aligned} \tag{7}$$

where $I(\widehat{\mathbf{H}}^{(r)}, T)$ indicates the mutual information between $\widehat{\mathbf{H}}^{(r)}$ and T, and $H(\widehat{\mathbf{H}}^{(r)}|T)$ indicates the entropy of $\widehat{\mathbf{H}}^{(r)}$ conditioned on T.

Based on Theorem 3.3, the common and private representations learned by our method are demonstrated to contain more task-relevant information and less task-irrelevant information than the node representations learned by previous contrastive learning methods. Note that we do not constrain the downstream task T as classification, regression, or clustering. As a result, the concatenated common and private representations learned by our method are expected to perform better on different downstream tasks.

3.4. Objective Function

Integrating the matching loss in Eq. (2), the correlation loss in Eq. (3), the reconstruction loss in Eq. (4), with the contrastive loss in Eq. (6), the objective function of the proposed DMG is formulated as:

$$\mathcal{J} = \mathcal{L}_{mat} + \alpha\mathcal{L}_{cor} + \beta\mathcal{L}_{rec} + \lambda\mathcal{L}_{con}, \tag{8}$$

where α, β and λ are non-negative parameters.

After optimization, the proposed DMG is expected to obtain complete and clean common representations, as well as more-complementarity and less-noise private representations, to achieve effective and robust UMGRL (verified in Section 4). We then conduct the average pooling (LeCun et al., 1989) to fuse private representations of all graphs to obtain the overall private representations $\mathbf{P}$, *i.e.,*

$$\mathbf{P} = \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \mathbf{P}^{(r)}. \tag{9}$$

Finally, we concatenate the overall private representations $\mathbf{P}$ with the common variable $\mathbf{S}$ to obtain final representations $\mathbf{Z}$. We list the pseudo-code of the proposed method in Appendix A.

4. Experiments

In this section, we conduct experiments on six public datasets to evaluate the proposed method in terms of different tasks. Details of experiments are shown in Appendix C, and additional results are shown in Appendix D.

4.1. Experimental Setup

4.1.1. Datasets

The used datasets include four multiplex graph datasets and two single-view graph datasets. Multiplex graph datasets include two citation datasets (*i.e.,* ACM (Wang et al., 2019) and DBLP (Wang et al., 2019)), two movie datasets (*i.e.,* IMDB (Wang et al., 2019) and Freebase (Wang et al., 2021)). Single-view graph datasets include two amazon sale datasets, *i.e.,* Photo and Computers (Shchur et al., 2018).

Table 1. Classification performance (*i.e.,* Macro-F1 and Micro-F1) of all methods on all multiplex graph datasets.

Method	ACM		IMDB		DBLP		Freebase	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Deep Walk	73.9 ± 0.3	74.1 ± 0.1	42.5 ± 0.2	43.3 ± 0.4	88.1 ± 0.2	89.5 ± 0.3	49.3 ± 0.3	52.1 ± 0.5
GCN	86.9 ± 0.2	87.0 ± 0.3	45.7 ± 0.4	49.8 ± 0.2	90.2 ± 0.2	90.9 ± 0.5	50.5 ± 0.2	53.3 ± 0.2
GAT	85.0 ± 0.4	84.9 ± 0.3	49.4 ± 0.2	53.6 ± 0.4	91.0 ± 0.4	92.1 ± 0.2	55.1 ± 0.3	59.7 ± 0.4
DGI	89.1 ± 0.4	88.2 ± 0.4	45.1 ± 0.2	46.7 ± 0.2	90.3 ± 0.1	91.1 ± 0.4	54.9 ± 0.1	58.2 ± 0.4
MNE	79.2 ± 0.4	79.7 ± 0.3	44.7 ± 0.5	45.6 ± 0.3	89.3 ± 0.2	90.6 ± 0.4	52.1 ± 0.3	54.3 ± 0.2
HAN	89.4 ± 0.2	89.2 ± 0.2	49.8 ± 0.5	54.2 ± 0.3	91.2 ± 0.4	92.0 ± 0.5	53.2 ± 0.1	57.2 ± 0.4
DMGI	89.8 ± 0.1	89.8 ± 0.1	52.2 ± 0.2	53.7 ± 0.3	92.1 ± 0.2	92.9 ± 0.3	54.9 ± 0.1	57.6 ± 0.3
DMGIattn	88.7 ± 0.3	88.7 ± 0.5	52.6 ± 0.2	53.6 ± 0.4	90.9 ± 0.2	91.8 ± 0.3	55.8 ± 0.4	58.3 ± 0.5
HDMI	90.1 ± 0.3	90.1 ± 0.3	55.6 ± 0.3	57.3 ± 0.3	91.3 ± 0.2	92.2 ± 0.5	56.1 ± 0.2	59.2 ± 0.2
HeCo	88.3 ± 0.3	88.2 ± 0.2	50.8 ± 0.3	51.7 ± 0.3	91.0 ± 0.3	91.6 ± 0.2	59.2 ± 0.3	61.7 ± 0.4
MCGC	90.2 ± 0.4	90.0 ± 0.3	56.3 ± 0.5	57.5 ± 0.6	91.9 ± 0.3	92.1 ± 0.4	56.6 ± 0.1	59.4 ± 0.3
CKD	90.4 ± 0.3	90.5 ± 0.2	54.8 ± 0.2	57.7 ± 0.3	92.0 ± 0.2	92.3 ± 0.5	60.4 ± 0.4	62.9 ± 0.5
DMG	**91.0 ± 0.3**	**90.9 ± 0.4**	**57.6 ± 0.2**	**58.9 ± 0.4**	**93.3 ± 0.2**	**94.0 ± 0.3**	**62.4 ± 0.7**	**65.9 ± 0.8**

4.1.2. Comparison Methods

The comparison methods include twelve single-view graph methods and eight multiplex graph methods. Single-view graph methods include two semi-supervised methods (GCN (Kipf & Welling, 2017) and GAT (Velickovic et al., 2018)), two traditional unsupervised methods (*i.e.,* DeepWalk (Perozzi et al., 2014) and VGAE (Kipf & Welling, 2016)), and eight self-supervised methods, (*i.e.,* DGI (Velickovic et al., 2019), GMI (Peng et al., 2020), MVGRL (Hassani & Khasahmadi, 2020), GRACE (Zhu et al., 2020b), GCA (Zhu et al., 2021)), GIC (Mavromatis & Karypis, 2021), COSTA (Zhang et al., 2022), and DSSL (Xiao et al., 2022)). Multiplex graph methods include one semi-supervised method (*i.e.,* HAN (Wang et al., 2019)), one traditional unsupervised method (*i.e.,* MNE (Zhang et al., 2018)), and six self-supervised methods, *i.e.,* DMGI (Park et al., 2020), DMGIattn (Park et al., 2020), HDMI (Jing et al., 2021a), HeCo (Wang et al., 2021), MCGC (Pan & Kang, 2021), and CKD (Zhou et al., 2022)).

For a fair comparison, we use single-view graph methods on multiplex graph datasets by separately learning the representations of each graph and further concatenating them for downstream tasks. Moreover, we apply random augmentations on every single-view graph dataset to generate two graph views for multiplex graph methods.

4.1.3. Evaluation Protocol

We follow previous works (Jing et al., 2021a; Zhou et al., 2022) to conduct node classification and node clustering as semi-supervised and unsupervised downstream tasks, respectively. Moreover, we employ Macro-F1 and Micro-F1 to evaluate the performance of node classification, and Accuracy and Normalized Mutual Information (NMI) to evaluate the performance of node clustering. Furthermore, we use noisy edges (*i.e.,* random edges) to randomly replace a certain ratio of edges in each graph, for evaluating the robustness of our method and comparison methods. The code is released at https://github.com/YujieMo/DMG.

4.2. Effectiveness Analysis

4.2.1. Effectiveness on the Multiplex Graph

We first evaluate the effectiveness of the proposed method on the multiplex graph datasets by reporting the results of node classification (*i.e.,* Macro-F1 and Micro-F1) and node clustering (*i.e.,* Accuracy and NMI) in Tables 1 and 2. Obviously, our method achieves the best effectiveness on both node classification task and node clustering task.

First, compared with single-view graph methods (*i.e.,* Deep Walk, GCN, GAT, and DGI), the proposed DMG always outperforms them by large margins. For example, the proposed DMG on average improves by 20.4%, compared to the best single-view graph method (*i.e.,* DGI), in terms of classification and clustering tasks, on all multiplex graph datasets. This demonstrates the superiority of the multiplex graph methods, which may explore correlations among different graphs and thus better mine the hidden information to learn discriminative node representations.

Second, compared to multiplex graph methods, the proposed DMG achieves the best results, followed by MCGC, CKD, HDMI, DMGI, HeCo, DMGIattn, HAN, and MNE. For example, our method on average improves by 1.6%, compared to the best comparison method MCGC, in terms of classification and clustering tasks, on all multiplex graph datasets. This can be attributed to the fact that the proposed DMG can explicitly capture complete and clean common information, as well as more-complementarity and less-noise private information. As a result, this introduces more task-relevant information and less task-irrelevant in learned representations, leading to better downstream task performance.

Table 2. Clustering performance (*i.e.,* Accuracy and NMI) of all methods on all multiplex graph datasets.

Method	ACM		IMDB		DBLP		Freebase	
	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI
Deep Walk	64.5 ± 0.7	41.6 ± 0.5	42.1 ± 0.4	1.5 ± 0.1	89.5 ± 0.4	69.0 ± 0.2	44.5 ± 0.6	12.8 ± 0.4
DGI	81.1 ± 0.6	64.0 ± 0.4	48.9 ± 0.2	8.3 ± 0.3	85.4 ± 0.3	65.6 ± 0.4	52.9 ± 0.2	17.8 ± 0.2
MNE	69.1 ± 0.2	54.5 ± 0.3	46.5 ± 0.3	4.6 ± 0.2	86.3 ± 0.3	68.4 ± 0.2	45.1 ± 0.5	13.3 ± 0.7
DMGI	88.4 ± 0.3	68.7 ± 0.5	52.5 ± 0.7	13.1 ± 0.3	91.8 ± 0.5	76.4 ± 0.6	53.1 ± 0.4	17.3 ± 0.4
DMGIattn	90.9 ± 0.4	70.2 ± 0.3	52.6 ± 0.3	9.2 ± 0.2	91.3 ± 0.4	75.2 ± 0.4	52.3 ± 0.5	17.1 ± 0.3
HDMI	90.8 ± 0.4	69.5 ± 0.5	57.6 ± 0.4	14.5 ± 0.4	90.1 ± 0.4	73.1 ± 0.3	58.3 ± 0.3	20.3 ± 0.4
HeCo	88.4 ± 0.6	67.8 ± 0.8	50.9 ± 0.5	10.1 ± 0.6	89.2 ± 0.3	71.0 ± 0.7	58.4 ± 0.6	20.4 ± 0.5
MCGC	90.4 ± 0.5	69.0 ± 0.5	56.5 ± 0.3	14.9 ± 0.4	91.9 ± 0.2	76.5 ± 0.4	58.1 ± 0.4	47.2 ± 0.3
CKD	90.6 ± 0.4	69.3 ± 0.3	53.9 ± 0.3	13.8 ± 0.4	91.4 ± 0.4	75.9 ± 0.4	58.5 ± 0.6	20.6 ± 0.4
DMG	**92.9 ± 0.3**	**74.5 ± 0.4**	**60.3 ± 0.5**	**17.0 ± 0.3**	**94.1 ± 0.4**	**80.0 ± 0.2**	**63.6 ± 0.6**	**21.8 ± 0.4**

Table 3. Classification and clustering performance (*i.e.,* Macro-F1, Micro-F1, Accuracy and NMI) on all single-view graph datasets.

Method	Photo				Computers			
	Macro-F1	Micro-F1	Accuracy	NMI	Macro-F1	Micro-F1	Accuracy	NMI
Deep Walk	87.4 ± 0.5	89.7 ± 0.3	46.2 ± 0.2	35.4 ± 0.3	84.0 ± 0.3	85.6 ± 0.4	32.5 ± 0.3	29.8 ± 0.2
VGAE	89.9 ± 0.2	91.6 ± 0.4	54.8 ± 0.5	37.4 ± 0.3	82.6 ± 0.3	85.3 ± 0.4	37.2 ± 0.3	32.4 ± 0.5
GCN	90.5 ± 0.3	92.5 ± 0.2	N/A	N/A	84.0 ± 0.4	86.4 ± 0.3	N/A	N/A
GAT	90.2 ± 0.5	91.8 ± 0.4	N/A	N/A	83.2 ± 0.2	85.7 ± 0.4	N/A	N/A
DGI	89.3 ± 0.2	91.6 ± 0.3	59.1 ± 0.4	43.2 ± 0.3	79.3 ± 0.3	83.9 ± 0.5	40.7 ± 0.3	33.4 ± 0.2
GMI	89.3 ± 0.4	90.6 ± 0.2	64.6 ± 0.2	47.2 ± 0.3	80.1 ± 0.4	82.2 ± 0.4	41.5 ± 0.2	34.5 ± 0.3
MVGRL	90.1 ± 0.3	91.7 ± 0.4	48.3 ± 0.5	34.4 ± 0.4	84.6 ± 0.6	86.9 ± 0.5	47.8 ± 0.5	47.1 ± 0.3
GRACE	90.3 ± 0.5	91.9 ± 0.3	74.1 ± 0.4	63.4 ± 0.2	84.2 ± 0.3	86.8 ± 0.5	49.6 ± 0.2	47.9 ± 0.4
GCA	91.1 ± 0.4	92.4 ± 0.4	73.6 ± 0.3	61.4 ± 0.2	85.9 ± 0.5	87.7 ± 0.3	51.3 ± 0.5	42.6 ± 0.4
GIC	90.0 ± 0.3	91.6 ± 0.2	69.5 ± 0.2	61.5 ± 0.1	82.6 ± 0.4	84.9 ± 0.3	52.5 ± 0.2	46.4 ± 0.2
COSTA	91.3 ± 0.4	92.5 ± 0.3	73.1 ± 0.3	62.1 ± 0.5	86.4 ± 0.3	**88.3 ± 0.4**	53.2 ± 0.2	48.1 ± 0.3
DSSL	90.6 ± 0.2	92.1 ± 0.3	74.5 ± 0.4	63.9 ± 0.5	85.6 ± 0.3	87.3 ± 0.4	53.5 ± 0.2	48.3 ± 0.4
DMG	**91.8 ± 0.2**	**92.9 ± 0.3**	**77.6 ± 0.3**	**66.5 ± 0.4**	**86.6 ± 0.3**	**88.3 ± 0.2**	**55.6 ± 0.4**	**49.3 ± 0.5**

"N/A" indicates that we did not evaluate semi-supervised methods (*i.e.,* GCN and GAT) on unsupervised tasks (*i.e.,* clustering).

4.2.2. Effectiveness on the single-view graph

To further verify the effectiveness of the proposed method on single-view graph datasets after random data augmentation, we report the results of node classification and node clustering on single-view graph datasets in Table 3. We can observe that our method achieves competitive results on both the node classification task and node clustering task.

First, compared to the semi-supervised baselines (*i.e.,* GCN and GAT), the proposed DMG obtains promising improvements. For example, the proposed DMG on average improves by 1.8%, compared to the best semi-supervised method (*i.e.,* GCN), in terms of the classification task, on all single-view graph datasets. Second, compared to all self-supervised methods (*i.e.,* DGI, GMI, MVGRL, GRACE, GCA, GIC, COSTA, and DSSL), the proposed DMG also achieves superior performance. For example, the proposed DMG on average outperforms the best self-supervised method (*i.e.,* DSSL) by 2.3%, in terms of classification and clustering tasks, on all single-view graph datasets. This indicates that on single-view graph datasets, the proposed DMG is still able to extract invariant common information between two augmented views, as well as preserve complementarity and remove noise in augmented graph structures. Therefore, the effectiveness of the proposed method is further validated on single-view graph datasets.

4.3. Robustness Analysis

We further evaluate the robustness of the proposed method on the multiplex dataset by reporting results of node classification and node clustering under different noisy edge ratios η in Figure 2.

From Figure 2, we have the observations as follows. First, compared to all self-supervised MGRL methods, the proposed DMG consistently achieves the best performance under different noise ratios on the DBLP dataset, demonstrating the superiority of the proposed method again. Second, with the increase of the noise ratios, the performance degradation of all self-supervised methods is much more drastically than the proposed method. For example, DMG and CKD achieve the Macro-F1 of 93.3 and 92.0 under

Table 4. Classification performance (*i.e.,* Macro-F1 and Micro-F1) of each component in the proposed method on all datasets.

$\mathcal{L}_{mat}$	$\mathcal{L}_{cor}$	$\mathcal{L}_{rec}$	$\mathcal{L}_{con}$	ACM		IMDB		DBLP		Freebase	
				Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
✓	–	–	✓	76.1 ± 0.5	75.8 ± 0.3	43.8 ± 0.4	45.5 ± 0.2	92.2 ± 0.4	93.0 ± 0.3	47.1 ± 0.6	47.5 ± 0.8
✓	–	✓	–	86.0 ± 0.3	86.0 ± 0.5	51.6 ± 0.5	54.0 ± 0.4	92.0 ± 0.2	92.9 ± 0.3	36.3 ± 0.6	40.2 ± 0.5
✓	✓	–	–	86.5 ± 0.4	86.1 ± 0.4	53.0 ± 0.3	54.7 ± 0.5	92.7 ± 0.3	93.5 ± 0.4	53.1 ± 0.8	55.1 ± 0.9
✓	–	✓	✓	74.3 ± 0.4	73.7 ± 0.6	46.0 ± 0.4	49.2 ± 0.5	91.7 ± 0.5	92.7 ± 0.3	35.7 ± 0.4	38.8 ± 0.7
✓	✓	–	✓	87.9 ± 0.6	88.0 ± 0.4	53.4 ± 0.3	56.1 ± 0.1	92.3 ± 0.4	93.0 ± 0.6	50.3 ± 0.7	52.9 ± 0.6
✓	✓	✓	–	87.8 ± 0.5	87.6 ± 0.3	54.7 ± 0.3	56.2 ± 0.5	92.5 ± 0.4	93.3 ± 0.6	55.1 ± 0.5	58.2 ± 0.7
✓	✓	✓	✓	$\mathbf{91.0 \pm 0.3}$	$\mathbf{90.9 \pm 0.4}$	$\mathbf{57.6 \pm 0.2}$	$\mathbf{58.9 \pm 0.4}$	$\mathbf{93.3 \pm 0.2}$	$\mathbf{94.0 \pm 0.3}$	$\mathbf{62.4 \pm 0.7}$	$\mathbf{65.9 \pm 0.8}$

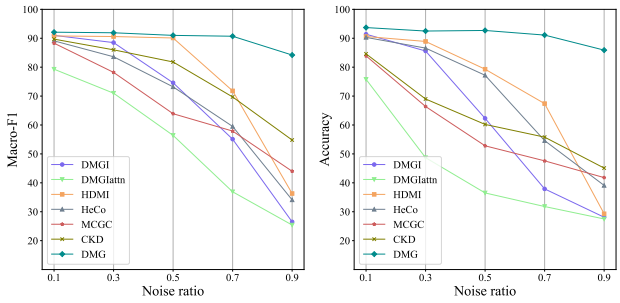


Figure 2. Classification and clustering performance of our method and all self-supervised MGRL methods under different noisy edges ratios η on the DBLP dataset.

$\eta = 0$, respectively, while with the increase of the noise rate, DMG is remarkably superior to CKD.

The reasons can be summarized as follows. On the one hand, our method extracts complete and clean common information through disentangled representation learning. As a result, the complete and clean common information is supposed to be free of noise. On the other hand, the contrastive constraint further preserves the complementarity and removes the noise in the private information. Therefore, the common and private representations learned by our method are expected to be robust to the noise in each graph.

4.4. Ablation Study

The proposed DMG investigates the matching loss, the correlation loss, and the reconstruction loss (*i.e.,* $\mathcal{L}_{mat}$, $\mathcal{L}_{cor}$, and $\mathcal{L}_{rec}$) to obtain disentangled common and private representations. Moreover, DMG further investigates the contrastive loss (*i.e.,* $\mathcal{L}_{con}$) to preserve the complementarity and remove the noise in private representations. To verify the effectiveness of each component of the objective function in the proposed method, we investigate the performance of all variants (except $\mathcal{L}_{mat}$ as we cannot obtain final representations without $\mathcal{L}_{mat}$) on the node classification task by reporting the results in Table 4.

According to Figure 4, we can draw the following conclusions. First, our method with the complete objective function achieves the best performance. For example, our method on average improves by 5.7%, compared to the best variant (*i.e.,* without $\mathcal{L}_{con}$), indicating that all the losses are necessary for our method. This is consistent with our above argument. That is, it is essential for UMGRL to consider complete and clean common information, as well as more-complementarity and less-noise private information. Second, the variant without $\mathcal{L}_{cor}$ performs significantly inferior to the other two variants (without $\mathcal{L}_{rec}$ and without $\mathcal{L}_{con}$, respectively). This makes sense as the correlation loss is needed to guarantee the independence between common and private representations, which is generally essential for disentangled representation learning.

5. Conclusion

In this paper, we proposed a disentangled representation learning framework for the multiplex graph. To do this, we first disentangled the common and private representations to capture complete and clean common information. We further designed a contrastive constraint to preserve the complementarity and remove the noise in the private information. Theoretical analysis indicates that the common and private representations learned by our method can be provably disentangled and contain more task-relevant information and less task-irrelevant information to benefit downstream tasks. Comprehensive experimental results demonstrate that the proposed method consistently outperforms state-of-the-art methods in terms of both effectiveness and robustness on different downstream tasks. We discuss potential limitations and future directions in Appendix E.

6. Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2022YFA1004100, and the Medico-Engineering Cooperation Funds from the University of Electronic Science and Technology of China under Grants No. ZYGX2022YGRH009 and ZYGX2022YGRH014.

References

Benton, A., Khayrallah, H., Gujral, B., Reisinger, D. A., Zhang, S., and Arora, R. Deep generalized canonical correlation analysis. In Augenstein, I., Gella, S., Ruder, S., Kann, K., Can, B., Welbl, J., Conneau, A., Ren, X., and Rei, M. (eds.), *Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019*, pp. 1–6, 2019.

Chen, B., Zhang, J., Zhang, X., Dong, Y., Song, J., Zhang, P., Xu, K., Kharlamov, E., and Tang, J. Gccad: Graph contrastive coding for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

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

Chen, X. and He, K. Exploring simple siamese representation learning. In *CVPR*, pp. 15750–15758, 2021.

Chu, X., Fan, X., Yao, D., Zhu, Z., Huang, J., and Bi, J. Cross-network embedding for multi-network alignment. In *WWW*, pp. 273–284, 2019.

Dong, Y., Chawla, N. V., and Swami, A. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*, pp. 135–144, 2017.

Donnat, C., Zitnik, M., Hallac, D., and Leskovec, J. Learning structural node embeddings via diffusion wavelets. In *KDD*, pp. 1320–1329, 2018.

Federici, M., Dutta, A., Forré, P., Kushman, N., and Akata, Z. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2020.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. Kernel constrained covariance for dependence measurement. In *AISTATES*, pp. 112–119, 2005.

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

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

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2016.

Jiang, W., Liu, S., Gao, C., Cao, J., He, R., Feng, J., and Yan, S. PSGAN: pose and expression robust spatial-aware GAN for customizable makeup transfer. In *CVPR*, pp. 5193–5201, 2020.

Jing, B., Park, C., and Tong, H. Hdmi: High-order deep multiplex infomax. In *WWW*, pp. 2414–2424, 2021a.

Jing, L., Vincent, P., LeCun, Y., and Tian, Y. Understanding dimensional collapse in contrastive self-supervised learning. In *ICLR*, 2021b.

Kingma, P. D. and Ba, L. J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kipf, N. T. and Welling, M. Semi-supervised classification with graph convolutional networks. In *ICLR*, pp. 1–14, 2017.

Kipf, T. N. and Welling, M. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. Handwritten digit recognition with a back-propagation network. In *NeurIPS*, volume 2, 1989.

Li, B., Jing, B., and Tong, H. Graph communal contrastive learning. In *WWW*, pp. 1203–1213, 2022.

Li, H., Wang, X., Zhang, Z., Yuan, Z., Li, H., and Zhu, W. Disentangled contrastive learning on graphs. In *NeurIPS*, volume 34, pp. 21872–21884, 2021.

Liu, Y., Tu, W., Zhou, S., Liu, X., Song, L., Yang, X., and Zhu, E. Deep graph clustering via dual correlation reduction. In *AAAI*, pp. 7603–7611, 2022.

Liu, Y., Yang, X., Zhou, S., Liu, X., Wang, S., Liang, K., Tu, W., and Li, L. Simple contrastive graph clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2023a.

Liu, Y., Yang, X., Zhou, S., Liu, X., Wang, Z., Liang, K., Tu, W., Li, L., Duan, J., and Chen, C. Hard sample aware network for contrastive deep graph clustering. In *AAAI*, 2023b.

Lyu, Q. and Fu, X. Nonlinear multiview analysis: Identifiability and neural network-assisted implementation. *IEEE Transactions on Signal Processing*, 68:2697–2712, 2020.

Lyu, Q., Fu, X., Wang, W., and Lu, S. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *ICLR*, 2022.

Ma, J., Cui, P., Kuang, K., Wang, X., and Zhu, W. Disentangled graph convolutional networks. In *ICML*, pp. 4212–4221, 2019.

Mavromatis, C. and Karypis, G. Graph infoclust: Maximizing coarse-grain mutual information in graphs. In *PAKDD*, pp. 541–553, 2021.

Mercatali, G., Freitas, A., and Garg, V. Symmetry-induced disentanglement on graphs. In *NeurIPS*, volume 35, pp. 31497–31511, 2022.

Mrabah, N., Bouguessa, M., Touati, M. F., and Ksantini, R. Rethinking graph auto-encoder models for attributed graph clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *ICML*, pp. 807–814, 2010.

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

Pan, E. and Kang, Z. Multi-view contrastive graph clustering. In *NeurIPS*, volume 34, pp. 2148–2159, 2021.

Park, C., Kim, D., Han, J., and Yu, H. Unsupervised attributed multiplex network embedding. In *AAAI*, pp. 5371–5378, 2020.

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

Perozzi, B., Al-Rfou, R., and Skiena, S. Deepwalk: Online learning of social representations. In *KDD*, pp. 701–710, 2014.

Shchur, O., Mumme, M., Bojchevski, A., and Günnemann, S. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

Shi, C., Hu, B., Zhao, W. X., and Yu, P. S. Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 31(2):357–370, 2019.

Shi, Y., Gui, H., Zhu, Q., Lance, K., and Han, J. Aspem: Embedding learning by aspects in heterogeneous information networks. In *SDM*, pp. 144–152, 2018a.

Shi, Y., Zhu, Q., Guo, F., Zhang, C., and Han, J. Easing embedding learning by comprehensive transcription of heterogeneous information networks. In *KDD*, pp. 2190–2199, 2018b.

Simonovsky, M. and Komodakis, N. Graphvae: Towards generation of small graphs using variational autoencoders. In *ICANN*, volume 11139, pp. 412–422, 2018.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *ICLR*, pp. 1–12, 2018.

Velickovic, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. Deep graph infomax. In *ICLR*, pp. 1–17, 2019.

Wang, H., Guo, X., Deng, Z.-H., and Lu, Y. Rethinking minimal sufficient representation in contrastive learning. In *CVPR*, pp. 16041–16050, 2022a.

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., and Yu, P. S. Heterogeneous graph attention network. In *WWW*, pp. 2022–2032, 2019.

Wang, X., Liu, N., Han, H., and Shi, C. Self-supervised heterogeneous graph neural network with co-contrastive learning. In *KDD*, pp. 1726–1736, 2021.

Wang, X., Chen, H., Tang, S., Wu, Z., and Zhu, W. Disentangled representation learning, 2022b.

Wang, X., Chen, H., Zhou, Y., Ma, J., and Zhu, W. Disentangled representation learning for recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):408–424, 2022c.

Xiao, T., Chen, Z., Guo, Z., Zhuang, Z., and Wang, S. Decoupled self-supervised learning for graphs. In *NeurIPS*, 2022.

Xie, D., Deng, C., Li, C., Liu, X., and Tao, D. Multi-task consistency-preserving adversarial hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 29:3626–3637, 2020.

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

Xu, J., Ren, Y., Tang, H., Pu, X., Zhu, X., Zeng, M., and He, L. Multi-VAE: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *ICCV*, pp. 9234–9243, 2021a.

Xu, J., Ren, Y., Tang, H., Yang, Z., Pan, L., Yang, Y., Pu, X., Yu, P. S., and He, L. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–12, 2022a.

Xu, J., Tang, H., Ren, Y., Peng, L., Zhu, X., and He, L. Multi-level feature learning for contrastive multi-view clustering. In *CVPR*, pp. 16051–16060, 2022b.

Xu, M., Wang, H., Ni, B., Guo, H., and Tang, J. Self-supervised graph-level representation learning with local and global structure. In *ICML*, volume 139, pp. 11548–11558, 2021b.

Yang, Y., Feng, Z., Song, M., and Wang, X. Factorizable graph convolutional networks. In *NeurIPS*, volume 33, pp. 20286–20296, 2020.

Zhang, C., Zhang, K., Zhang, C., Pham, T. X., Yoo, C. D., and Kweon, I. S. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. In *ICLR*, 2021.

Zhang, H. and Kou, G. Role-based multiplex network embedding. In *ICML*, pp. 26265–26280, 2022.

Zhang, H., Qiu, L., Yi, L., and Song, Y. Scalable multiplex network embedding. In *IJCAI*, pp. 3082–3088, 2018.

Zhang, Y., Zhu, H., Song, Z., Koniusz, P., and King, I. Costa: Covariance-preserving feature augmentation for graph contrastive learning. In *KDD*, pp. 2524–2534, 2022.

Zhou, S., Yu, K., Chen, D., Li, B., Feng, Y., and Chen, C. Collaborative knowledge distillation for heterogeneous information network embedding. In *WWW*, pp. 1631–1639, 2022.

Zhu, Y., Min, M. R., Kadav, A., and Graf, H. P. S3VAE: self-supervised sequential VAE for representation disentanglement and data generation. In *CVPR*, pp. 6537–6546, 2020a.

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

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., and Wang, L. Graph contrastive learning with adaptive augmentation. In *WWW*, pp. 2069–2080, 2021.

Zhu, Y., Xu, Y., Cui, H., Yang, C., Liu, Q., and Wu, S. Structure-enhanced heterogeneous graph contrastive learning. In *SDM*, pp. 82–90, 2022.

A. Algorithm

This section provides the pseudo-code of the proposed method.

Algorithm 1 The pseudo-code of the proposed DMG.

Input: Node features $\mathbf{X}^{(r)}$ and graph structure $\mathbf{A}^{(r)}$ of graph $\mathbf{G}^{(r)}$ for $\forall r \in [1, \mathcal{R}]$, non-negative parameters α, β and λ, maximum training steps E;
Output: Encoders $f_c^{(r)} \circ g^{(r)}$, $f_p^{(r)} \circ g^{(r)}$, decoder $\boldsymbol{p}^{(r)}$, and measurable functions $\phi^{(r)}$ and $\psi^{(r)}$;
Initialize parameters;
while not reaching E **do**
  Obtain common variable $\mathbf{S}$ with orthogonality and zero mean via singular value decomposition;
  while not reaching E **do**
    Obtain common and private representations (*i.e.,* $\mathbf{C}^{(r)}$ and $\mathbf{P}^{(r)}$) with encoders $f_c^{(r)} \circ g^{(r)}$, $f_p^{(r)} \circ g^{(r)}$;
    Conduct the matching loss between $\mathbf{C}^{(r)}$ and $\mathbf{S}$ by Eq. (2);
    Conduct the correlation loss between $\mathbf{C}^{(r)}$ and $\mathbf{P}^{(r)}$ under measurable functions $\phi^{(r)}$ and $\psi^{(r)}$ by Eq. (3);
    Conduct the reconstruction loss between reconstructed node representations and original input by Eq. (4);
    Calculate the cosine similarity $\epsilon_{ij}^{(r)}$ between $\mathbf{s}_i$ and $\mathbf{s}_j$ of the node pair $(v_i, v_j) \in \mathcal{E}^{(r)}$ by Eq. (5);
    Conduct the contrastive loss on $\mathbf{P}^{(r)}$ based on complementary and noisy edges set $\mathcal{E}_c^{(r)}$ and $\mathcal{E}_p^{(r)}$ by Eq. (6);
    Compute the objective function $\mathcal{J}$ by Eq. (8);
    Back-propagate $\mathcal{J}$ to update model weights;
  end while
end while

B. Proofs in Section 3

This section provides detailed proofs of Theorems in Section 3.

B.1. Proof of Theorem 3.1

Proof. Consider the matching loss in Eq. (2) achieves its minimization, which indicates that the common representations $\mathbf{C}^{(r)}$ from different graphs are perfectly aligned, *i.e.,* $\mathbf{C}^{(r)} = \mathbf{C}^{(r')}$ ($r \neq r'$). Assuming that the solution $f_c^{(r)}$ satisfying the constraint in Eq. (2) has been found, then we have:

$$f_c^{(r)} \circ g^{(r)}(\mathbf{X}^{(r)}, \mathbf{A}^{(r)}) = f_c^{(r')} \circ g^{(r')}(\mathbf{X}^{(r')}, \mathbf{A}^{(r')}). \tag{10}$$

As the solution satisfies the correlation loss in Eq. (3) and the reconstruction loss in Eq. (4), the common and private representations (*i.e.,* $\mathbf{C}^{(r)}$ and $\mathbf{P}^{(r)}$) are expected to be statistical independent, and $f_c^{(r)} \circ g^{(r)}$ are expected to be invertible. We simply use $q_c^{(r)}$ to denote the inverted function $g^{(r)^{(-1)}} \circ f_c^{(r)^{(-1)}}$. Denote the optimal common and private representations as $\mathbf{C}_*$ and $\mathbf{P}_*^{(r)}$, which contain complete and clean common and private information, respectively. Note that the optimal common and private representations are also statistically independent. According to the invertibility of $q_c^{(r)}$ and the independence between $\mathbf{C}_*$ and $\mathbf{P}_*^{(r)}$, we can transform Eq. (10) to:

$$q_c^{(r)}\left(\begin{bmatrix} \mathbf{C}_* \\ \mathbf{P}_*^{(r)} \end{bmatrix}\right) = q_c^{(r')}\left(\begin{bmatrix} \mathbf{C}_* \\ \mathbf{P}_*^{(r')} \end{bmatrix}\right). \tag{11}$$

where $q_c^{(r)}(\vartheta^{(r)}) = g^{(r)^{(-1)}} \circ f_c^{(r)^{(-1)}}(\vartheta^{(r)})$, and $\vartheta^{(r)} = [\mathbf{C}_*^\top, (\mathbf{P}_*^{(r)})^\top]^\top$. Therefore, to demonstrate the functions $f_c^{(r)}$ can extract complete and clean common information, we only have to demonstrate that $q_c^{(r)}$ is the function of only $\mathbf{C}_*$ but not the function of $\mathbf{P}_*^{(r)}$. To do this, we then calculate the Jacobian of $q^{(r)}$ to analyze the first-order partial derivatives of $q_c^{(r)}$ and $q_p^{(r)}$ w.r.t. $\mathbf{C}_*$ and $\mathbf{P}_*^{(r)}$. The Jacobian of $q^{(r)}$ can be formulated as:

$$\boldsymbol{J}^{(r)} = \begin{bmatrix} \boldsymbol{J}_{11}^{(r)} & \boldsymbol{J}_{12}^{(r)} \\ \boldsymbol{J}_{21}^{(r)} & \boldsymbol{J}_{22}^{(r)} \end{bmatrix}, \tag{12}$$

where $\boldsymbol{J}_{11}^{(r)} \in \mathbb{R}^{D\times D}$, $\boldsymbol{J}_{12}^{(r)} \in \mathbb{R}^{D\times d}$, $\boldsymbol{J}_{21}^{(r)} \in \mathbb{R}^{d\times D}$ and $\boldsymbol{J}_{22}^{(r)} \in \mathbb{R}^{d\times d}$ are Jacobian matrices, and elements of them can be formulated as:

$$\begin{aligned}\left[\boldsymbol{J}_{11}^{(r)}\right]_{i,j} &= \frac{\partial\left[q_c^{(r)}(\boldsymbol{\vartheta}^{(r)})\right]_i}{\partial \mathbf{c}_{*j}^{(r)}}, \left[\boldsymbol{J}_{12}^{(r)}\right]_{i,k} = \frac{\partial\left[q_c^{(r)}(\boldsymbol{\vartheta}^{(r)})\right]_i}{\partial \mathbf{p}_{*k}^{(r)}},\\ \left[\boldsymbol{J}_{21}^{(r)}\right]_{k,i} &= \frac{\partial\left[q_p^{(r)}(\boldsymbol{\vartheta}^{(r)})\right]_k}{\partial \mathbf{c}_{*i}^{(r)}}, \left[\boldsymbol{J}_{22}^{(r)}\right]_{k,l} = \frac{\partial\left[q_p^{(r)}(\boldsymbol{\vartheta}^{(r)})\right]_k}{\partial \mathbf{p}_{*l}^{(r)}},\end{aligned} \tag{13}$$

where $i, j \in [1, D], k, l \in [1, d]$. Then we only have to demonstrate that $\boldsymbol{J}_{12}^{(r)}$ is an all-zero matrix while the determinant of $\boldsymbol{J}_{11}^{(r)}$ is non-zero to show that the matrix consisting of all the partial derivatives of $q_c^{(r)}$ w.r.t. $\mathbf{C}_*$ is full rank while any partial derivatives of $q_c^{(r)}$ w.r.t. $\mathbf{P}_*^{(r)}$ is zero.

Note that Eq. (11) holds over the whole latent space. Therefore, with any fixed $\bar{\mathbf{C}}_*$ and $\bar{\mathbf{P}}_*^{(r')}$, for all $\mathbf{P}_*^{(r)}$, we have:

$$q_c^{(r)}\left(\begin{bmatrix}\bar{\mathbf{C}}_*\\ \mathbf{P}_*^{(r)}\end{bmatrix}\right) = q_c^{(r')}\left(\begin{bmatrix}\bar{\mathbf{C}}_*\\ \bar{\mathbf{P}}_*^{(r')}\end{bmatrix}\right). \tag{14}$$

Then we take the partial derivatives of Eq. (14) w.r.t. $\mathbf{p}_j^{(r)}$ for $j \in [1, d]$, and we have: $\boldsymbol{J}_{12}^{(r)}|_{\bar{\mathbf{C}},\mathbf{P}^{(r)}} = \boldsymbol{J}_{12}^{(r')}|_{\bar{\mathbf{C}},\bar{\mathbf{P}}^{(r')}}$. According to the chain rules and taking derivatives of constants, we further have:

$$\boldsymbol{J}_{12}^{(r')}|_{\bar{\mathbf{C}},\bar{\mathbf{P}}^{(r')}} = \left(\boldsymbol{J}_{q_c^{(r')}}\Big|_{\bar{\mathbf{C}},\bar{\mathbf{P}}^{(r')}}\right)\begin{bmatrix}\mathbf{0}_{D\times d}\\ \mathbf{0}_{D\times d}\end{bmatrix} = \mathbf{0}_{D\times d}, \tag{15}$$

where $\boldsymbol{J}_{q_c^{(r')}} \in \mathbb{R}^{D\times(D+d)}$ is the Jacobian of $q_c^{(r')}$. Note that the equation above holds for any fixed $\bar{\mathbf{C}}_*$ and $\bar{\mathbf{P}}_*^{(r)}$, and thus the same derivation holds for all $\mathbf{C}_*$ and $\mathbf{P}_*^{(r)}$. Therefore, $\boldsymbol{J}_{12}^{(r)}$ is an all-zero matrix and the learned $q_c^{(r)}(\vartheta^{(r)})$ is not a function of $\mathbf{P}_*^{(r)}$.

Based on the above proof, we can rewrite Eq. (12) as:

$$\boldsymbol{J}^{(r)} = \begin{bmatrix}\boldsymbol{J}_{11}^{(r)} & \mathbf{0}_{D\times d}\\ \boldsymbol{J}_{21}^{(r)} & \boldsymbol{J}_{22}^{(r)}\end{bmatrix}. \tag{16}$$

According to the property of determinant of block matrix and the invertibility of $q_c^{(r)}$, we have:

$$\det(\boldsymbol{J}^{(r)}) = \det(\boldsymbol{J}_{11}^{(r)})\det(\boldsymbol{J}_{22}^{(r)}) \neq 0. \tag{17}$$

This indicates that $\det(\boldsymbol{J}_{11}^{(r)}) \neq 0$ and $\det(\boldsymbol{J}_{22}^{(r)}) \neq 0$. Therefore, $\boldsymbol{J}_{11}^{(r)}$ is an non-zero matrix and $q_c^{(r)}$ is the function of only $\mathbf{C}_*$ but not the function of $\mathbf{P}_*^{(r)}$, *i.e.,* for $\forall r \in [1, \mathcal{R}]$, we have $\mathbf{C}^{(r)} = \varphi(\mathbf{C}_*)$, where φ is an invertible function as $\det(\boldsymbol{J}_{11}^{(r)}) \neq 0$. Therefore we complete the proof. □

B.2. Proof of Theorem 3.3

In the following proofs, for random variables A, B, C, we use $I(A, B)$ to represent the mutual information between A and B, and we use $I(A, B|C)$ to represent conditional mutual information of A and B on a given C, use $H(A)$ for the entropy, and $H(A|B)$ for the conditional entropy. We first list some properties of mutual information and entropy that will be used in the proofs.

- **Property 1.** Relationship between the mutual information and entropy:

$$I(A, B) = H(A) - H(A \mid B). \tag{18}$$

- **Property 2.** Relationship between the conditional mutual information and entropy:

$$I(A, B \mid C) = H(A \mid C) - H(A \mid B, C). \tag{19}$$

- **Property 3.** Chain rule of the mutual information:

$$I(A, B \mid C) = I(A \mid B) - I(A, B, C). \tag{20}$$

- **Property 4.** Relationship between the conditional entropy and entropy:

$$H(A \mid B) = H(A, B) - H(B). \tag{21}$$

To aid in the proof of Theorem 3.3, we first have the following Lemma:

Lemma B.1. *Given a downstream task T, the node representations $\widehat{\mathbf{H}}^{(r)}$ learned by our method, and the node representations $\widetilde{\mathbf{H}}^{(r)}$ learned by previous contrastive learning methods, which maximize the mutual information among different graphs (*i.e., $I(\widetilde{\mathbf{H}}^{(r)}, \widetilde{\mathbf{H}}^{(r')})$*), we have*

$$I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}, T) = I(\widetilde{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}, T) = I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')}, T), \tag{22}$$

$$H(\widehat{\mathbf{H}}^{(r)}) - H(\widetilde{\mathbf{H}}^{(r)}) = H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}) - H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}). \tag{23}$$

Proof. According to the proof in Theorem 3.1, our method is available to obtain the complete and clean common information among different graphs, then we have

$$I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}) = I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')}). \tag{24}$$

Assume previous contrastive learning methods also obtain complete information among different graphs via mutual information maximization, then we have

$$I(\widetilde{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}) = I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')}). \tag{25}$$

We then introduce an assumption, which is widely used in previous works (Federici et al., 2020; Wang et al., 2022a), *i.e.,* if the random variable C is observed, then random variable A is conditionally independent from any other variable B, *i.e.,* $I(A, B|C) = 0, \forall B$. Based on Eq. (24), Eq. (25), Properties 2-3, and the assumption above, we have

$$\begin{aligned}
&I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')}, T) - I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}, T) \\
&= [I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')}) - I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')} \mid T)] - [I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}) - I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')} \mid T)] \\
&= I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')} \mid T) - I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')} \mid T) \\
&= [H(\mathcal{G}^{(r')} \mid T) - H(\mathcal{G}^{(r')} \mid \widehat{\mathbf{H}}^{(r)}, T)] - [H(\mathcal{G}^{(r')} \mid T) - H(\mathcal{G}^{(r')}|\mathcal{G}^{(r)}, T)] \\
&= H(\mathcal{G}^{(r')} \mid \mathcal{G}^{(r)}, T) - H(\mathcal{G}^{(r')} \mid \widehat{\mathbf{H}}^{(r)}, T) \\
&= [I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')} \mid \mathcal{G}^{(r)}, T) + H(\mathcal{G}^{(r')} \mid \mathcal{G}^{(r)}, \widehat{\mathbf{H}}^{(r)}, T)] \\
&\quad - [I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')} \mid \widehat{\mathbf{H}}^{(r)}, T) + H(\mathcal{G}^{(r')} \mid \mathcal{G}^{(r)}, \widehat{\mathbf{H}}^{(r)}, T)] \\
&= I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')} \mid \mathcal{G}^{(r)}, T) - I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')} \mid \widehat{\mathbf{H}}^{(r)}, T) \\
&= I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')} \mid \mathcal{G}^{(r)}, T) \\
&= 0.
\end{aligned} \tag{26}$$

Similarly, we can obtain $I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')}, T) - I(\widetilde{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}, T) = 0$. Therefore, we have $I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}, T) = I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')}, T) = I(\widetilde{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}, T)$.

In addition, based on Eq. (24), Eq. (25), Properties 1 and 4, we further have

$$\begin{aligned}
&H(\widehat{\mathbf{H}}^{(r)}) - H(\widetilde{\mathbf{H}}^{(r)}) - H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}) + H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}) \\
&= H(\widehat{\mathbf{H}}^{(r)}) - H(\widetilde{\mathbf{H}}^{(r)}) - H(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}) + H(\mathcal{G}^{(r')}) + H(\widetilde{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}) - H(\mathcal{G}^{(r')}) \\
&= H(\widehat{\mathbf{H}}^{(r)}) - H(\widetilde{\mathbf{H}}^{(r)}) - H(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}) + H(\widetilde{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}) \\
&= H(\widehat{\mathbf{H}}^{(r)}) - H(\widetilde{\mathbf{H}}^{(r)}) - H(\widehat{\mathbf{H}}^{(r)}) + H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}) + \widetilde{\mathbf{H}}^{(r)} - H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}) \\
&= H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}) - H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}) \\
&= H(\widehat{\mathbf{H}}^{(r)}) - I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}) - H(\widehat{\mathbf{H}}^{(r)}) + I(\widetilde{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}) \\
&= 0.
\end{aligned} \tag{27}$$

Therefore, we have $H(\widehat{\mathbf{H}}^{(r)}) - H(\widetilde{\mathbf{H}}^{(r)}) = H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}) - H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')})$ and we complete the proof. □

Now we can prove the Theorem 3.3.

Proof of Theorem 3.3. We divide the proof into two parts, *i.e.,* 1) $I(\widehat{\mathbf{H}}^{(r)}, T) \geq I(\widetilde{\mathbf{H}}^{(r)}, T)$ and 2) $H(\widehat{\mathbf{H}}^{(r)}|T) \leq H(\widetilde{\mathbf{H}}^{(r)}|T)$. We first prove that $I(\widehat{\mathbf{H}}^{(r)}, T) \geq I(\widetilde{\mathbf{H}}^{(r)}, T)$ holds. Denote the complementary information of the representations learned by our method and previous contrastive learning method as $I(\widehat{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')})$ and $I(\widetilde{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')})$, respectively. With the contrastive loss in Eq. (6) achieving its minimum and thus preserving the complementarity in each graph, we have $I(\widehat{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}) \geq I(\widetilde{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')})$. Note that with Property 3, we have

$$I(\widehat{\mathbf{H}}^{(r)}, T) = I(\widehat{\mathbf{H}}^{(r)}, T, \mathcal{G}^{(r')}) + I(\widehat{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}). \tag{28}$$

According to Eq. (22) in Lemma B.1, *i.e.,* $I(\widehat{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}, T) = I(\widetilde{\mathbf{H}}^{(r)}, \mathcal{G}^{(r')}, T) = I(\mathcal{G}^{(r)}, \mathcal{G}^{(r')}, T)$, then we have

$$\begin{aligned}
I(\widehat{\mathbf{H}}^{(r)}, T) &= I(\widetilde{\mathbf{H}}^{(r)}, T, \mathcal{G}^{(r')}) + I(\widehat{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}) \\
&= I(\widetilde{\mathbf{H}}^{(r)}, T) - I(\widetilde{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}) + I(\widehat{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}).
\end{aligned} \tag{29}$$

Based on $I(\widehat{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}) \geq I(\widetilde{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')})$, we have $I(\widehat{\mathbf{H}}^{(r)}, T) \geq I(\widetilde{\mathbf{H}}^{(r)}, T)$.

Similar to the above, we denote the noisy information of the representations learned by our method and previous contrastive learning method as $H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}, T)$ and $H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}, T)$, respectively. With the contrastive loss in Eq. (6) achieving its minimum and thus removing the noise in each graph, we have $H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}, T) \leq H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}, T)$. According to Properties 1-3, and Eq. (23) in Lemma B.1, *i.e.,* $H(\widehat{\mathbf{H}}^{(r)}) - H(\widetilde{\mathbf{H}}^{(r)}) = H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}) - H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')})$, then we have

$$\begin{aligned}
H(\widehat{\mathbf{H}}^{(r)}|T) &= H(\widehat{\mathbf{H}}^{(r)}) - I(\widehat{\mathbf{H}}^{(r)}, T) \\
&= H(\widehat{\mathbf{H}}^{(r)}) - [I(\widehat{\mathbf{H}}^{(r)}, T, \mathcal{G}^{(r')}) + I(\widehat{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')})] \\
&= H(\widehat{\mathbf{H}}^{(r)}) - I(\widetilde{\mathbf{H}}^{(r)}, T, \mathcal{G}^{(r')}) - I(\widehat{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}) \\
&= H(\widehat{\mathbf{H}}^{(r)}) - I(\widetilde{\mathbf{H}}^{(r)}, T) + I(\widetilde{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}) - I(\widehat{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}) \\
&= H(\widehat{\mathbf{H}}^{(r)}) - [H(\widetilde{\mathbf{H}}^{(r)}) - H(\widetilde{\mathbf{H}}^{(r)}|T)] + I(\widetilde{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}) - I(\widehat{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}) \\
&= H(\widetilde{\mathbf{H}}^{(r)}|T) + H(\widehat{\mathbf{H}}^{(r)}) - H(\widetilde{\mathbf{H}}^{(r)}) + I(\widetilde{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}) - I(\widehat{\mathbf{H}}^{(r)}, T|\mathcal{G}^{(r')}) \\
&= H(\widetilde{\mathbf{H}}^{(r)}|T) + H(\widehat{\mathbf{H}}^{(r)}) - H(\widetilde{\mathbf{H}}^{(r)}) + H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}) \\
&\quad - H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}, T) - H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}) + H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}, T) \\
&= H(\widetilde{\mathbf{H}}^{(r)}|T) - H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}, T) + H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}, T)
\end{aligned} \tag{30}$$

Based on $H(\widehat{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}, T) \leq H(\widetilde{\mathbf{H}}^{(r)}|\mathcal{G}^{(r')}, T)$, we have $H(\widehat{\mathbf{H}}^{(r)}|T) \leq H(\widetilde{\mathbf{H}}^{(r)}|T)$. Therefore, we complete the proof. □

C. Experimental Settings

This section provides detailed experimental settings in Section 4, including the description of all datasets in Section C.1, summarization of all comparison methods in Section C.2, model architectures and settings in Section C.3, and the evaluation protocol in Section C.4.

Table 5. Statistics of all datasets.

Datasets	Nodes	Meta-paths	Edges	Features	Labeled Nodes	Classes
ACM	3,025	Paper-Subject-Paper (PSP) Paper-Author-Paper (PAP)	2,210,761 29,281	1,830 (Paper Abstract)	600	3
IMDB	4,780	Movie-Actor-Movie (MAM) Movie-Director-Movie (MDM)	98,010 21,018	1,232 (Movie Plot)	300	3
DBLP	4,057	Author-Paper-Author (APA) Author-Paper-Conference-Paper-Author (APCPA) Author-Paper-Term-Paper-Author (APTPA)	11,113 5,000,495 6,776,335	334 (Paper Abstract)	800	4
Freebase	3,492	Movie-Actor-Movie (MAM) Movie-Director-Movie (MDM) Movie-Writer-Movie (MWM)	254,702 8,404 10,706	3,492 (One-hot Encoding)	60	3
Photo	7,487	Product-Customer-Product (PCP)	287,326	745 (Product Reviews)	765	8
Computers	13,752	Product-Customer-Product (PCP)	574,418	767 (Product Reviews)	1375	10

C.1. Datasets

We use four public multiplex graph datasets and two single-view graph datasets from various domains. Multiplex graph datasets include two citation multiplex graph datasets (*i.e.,* ACM (Wang et al., 2019) and DBLP (Wang et al., 2019)), two movie multiplex graph datasets (*i.e.,* IMDB (Wang et al., 2019) and Freebase (Wang et al., 2021)). Single-view graph datasets include two amazon sale datasets (*i.e.,* Photo and Computers (Shchur et al., 2018)). Table 5 summarizes the data statistics. We list the details of the datasets as follows.

- **ACM**[1] contains 3,025 papers with graphs generated by two meta-paths (*i.e.,* paper-author-paper and paper-subject-paper). The feature of each paper is a 1,830-dimensional bag-of-words representation of the abstract. Papers are categorized into three classes, *i.e.,* database, wireless communication, and data mining.
- **IMDB**[2] contains 4,780 movies with graphs generated by two meta-paths (*i.e.,* movie-actor-movie and movie-director-movie). The feature of each movie is a 1,232-dimensional bag-of-words representation of its plots. Movies are categorized into three classes, *i.e.,* action, comedy, and drama.
- **DBLP**[3] contains 4,057 papers with graphs generated by three meta-paths (*i.e.,* author-paper-author, author-paper-conference-paper-author, and author-paper-term-paper-author). The feature of each paper is a 334-dimensional bag-of-words representation of its abstracts. Papers are categorized into four classes, *i.e.,* database, data mining, machine learning, and information retrieval.
- **Freebase**[4] contains 3,492 movies with graphs generated by three meta-paths (*i.e.,* movie-actor-movie, movie-director-movie and movie-writer-movie). We assign one-hot encoding to this dataset as no features are provided. Movies are categorized into four classes, ie action, comedy and drama.
- **Photo** and **Computers**[5] contain 7,487 and 13,752 products, respectively. Edges in each dataset indicate that two products are frequently bought together. The feature of each product is bag-of-words encoded product reviews. Products are categorized into several classes by the product category.

[1]https://www.acm.org/
[2]https://www.imdb.org/
[3]https://aminer.org/AMinerNetwork/
[4]http://www.freebase.com/
[5]https://docs.dgl.ai/en/0.6.x/api/python/dgl.data.html

Table 6. The characteristics of all comparison methods.

Methods	Multiplex	Single-view	Semi-sup	Tra-unsup	Self-sup	Features	Com-info	Pri-info	Comp-noise
GCN		✓	✓			✓			
GAT		✓	✓			✓			
DeepWalk		✓		✓					
VGAE		✓		✓		✓			
DGI		✓			✓	✓			
GMI		✓			✓	✓			
MVGRL		✓			✓	✓	✓		
GRACE		✓			✓	✓	✓		
GCA		✓			✓	✓	✓		
GIC		✓			✓	✓			
COSTA		✓			✓	✓	✓		
DSSL		✓			✓	✓	✓	✓	
MNE	✓			✓					
HAN	✓		✓			✓			
DMGI	✓				✓	✓			
DMGIattn	✓				✓	✓			
HDMI	✓				✓	✓			
HeCo	✓				✓	✓	✓		
MCGC	✓				✓	✓	✓		
CKD	✓				✓	✓	✓		
DMG (ours)	✓					✓	✓	✓	✓

C.2. Comparison Methods

The comparison methods include twelve methods designed for the single-view graph and eight for the multiplex graph, *i.e.,* GCN (Kipf & Welling, 2017), GAT (Velickovic et al., 2018), DeepWalk (Perozzi et al., 2014), VGAE (Kipf & Welling, 2016), DGI (Velickovic et al., 2019), GMI (Peng et al., 2020), MVGRL (Hassani & Khasahmadi, 2020), GRACE (Zhu et al., 2020b), GCA (Zhu et al., 2021), GIC (Mavromatis & Karypis, 2021), COSTA (Zhang et al., 2022), DSSL (Xiao et al., 2022), HAN (Wang et al., 2019), MNE (Zhang et al., 2018), DMGI (Park et al., 2020), DMGIattn (Park et al., 2020), HDMI (Jing et al., 2021a), HeCo (Wang et al., 2021), MCGC (Pan & Kang, 2021), and CKD (Zhou et al., 2022). The characteristics of all methods are listed in Table 6, where "Multiplex" and "Single-view" indicate the methods designed for the multiplex graph and single-view graph, respectively. "Semi-sup", "Tra-unsup", and "Self-sup" indicate that the method conducts semi-supervised learning, traditional unsupervised learning, and self-supervised learning, respectively. Note that our DMG is a new unsupervised framework that cannot be simply classified as traditional unsupervised learning or self-supervised learning. "Features" indicates that the method takes the node features into account. "Com-info", "Pri-info", and "Comp-noise" indicate that the method takes the common information, private information, complementarity and noise into account, respectively.

C.3. Model Architectures and Settings

As described in Section 3, the proposed DMG employs the one-layer GCN and MLP as the encoders (*i.e.,* $g^{(r)}$ and $f^{(r)}$) to obtain common representations $\mathbf{C}^{(r)} \in \mathbb{R}^{N\times D}$, and private representations $\mathbf{P}^{(r)} \in \mathbb{R}^{N\times d}$. Note that we assign different weights ω to the self-connection of multiplex graph datasets and single-view graph datasets during the graph convolution. Then the proposed DMG investigates the correlation loss to enforce the independence between common and private representations with the measurable functions (*i.e.,* $\phi^{(r)}$ and $\psi^{(r)}$). In the proposed DMG, the measurable functions $\phi^{(r)}$ and $\psi^{(r)}$ are implemented by the two-layer MLP. After that, the proposed DMG investigates the reconstruction loss to promote the invertibility of encoders with the reconstruction network $\boldsymbol{p}^{(r)}$, which is also implemented as an MLP. In the proposed DMG, all parameters were optimized by the Adam optimizer (Kingma & Ba, 2015) with initial learning rate and weight decay (1e-3 and 1e-4, respectively). We apply the ReLU function (Nair & Hinton, 2010) as a nonlinear activation function. In all experiments, we repeat the experiments five times for all methods and report the average results. Table 7 describes the detailed settings and architecture for most of our experimental setups with DMG.

Table 7. Settings for the proposed DMG.

Settings	ACM	IMDB	DBLP	Freebase	Photo	Computers
D	8	8	8	8	16	40
d	2	2	2	2	2	4
ω	3	3	3	3	1	1
Hidden units of $g^{(r)}$	256	512	256	256	256	512
Hidden units of $\phi^{(r)}$	256	256	256	256	256	256
Hidden units of $\psi^{(r)}$	256	256	256	256	256	256
Layers of $\boldsymbol{p}^{(r)}$	3	2	2	2	2	2
Learning rate	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3
Weight decay	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
Dropout	0.1	0.1	0.1	0.1	0.1	0.1

C.4. Evaluation Protocol

We follow the evaluation in previous works (Jing et al., 2021a; Pan & Kang, 2021; Zhou et al., 2022), where the model is trained in an unsupervised manner. Then, the learned representations are evaluated by several downstream tasks (*i.e.,* node classification and node clustering). For the node classification task, we evaluate the effectiveness of all methods with Micro-F1 and Macro-F1 scores. For the node clustering task, we evaluate the effectiveness of all methods with Accuracy score and Normalized Mutual Information (NMI) score, and NMI $= 2I(\hat{Y};Y)/[H(\hat{Y})+H(Y)]$, where $\hat{Y}$ and Y refer to the predicted cluster indexes and class labels.

D. Additional Experiments

This section provides some additional experimental results to further support the proposed method, including parameter analysis in Section D.1, additional ablation study in Section D.2, visualization of disentangled representation learning in Section D.3, comparison experiments with hard matching loss and InfoNCE loss in Section D.4, and additional results of the robustness on other multiplex datasets in Figures 6 and 7.

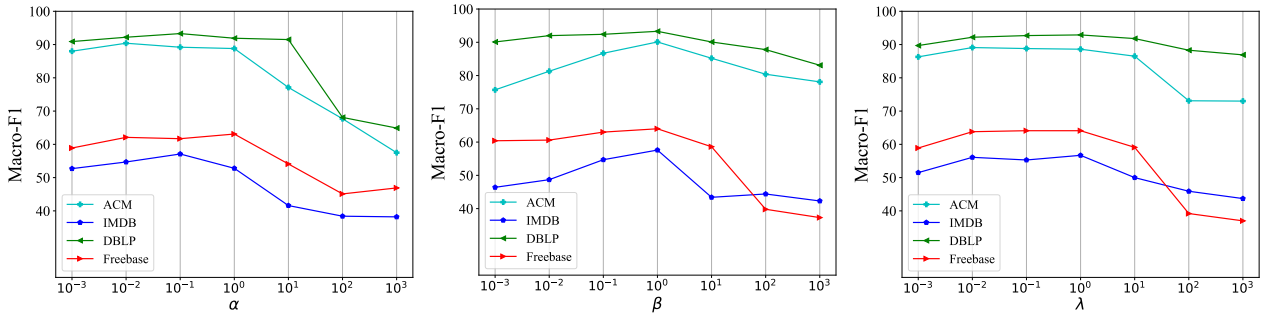
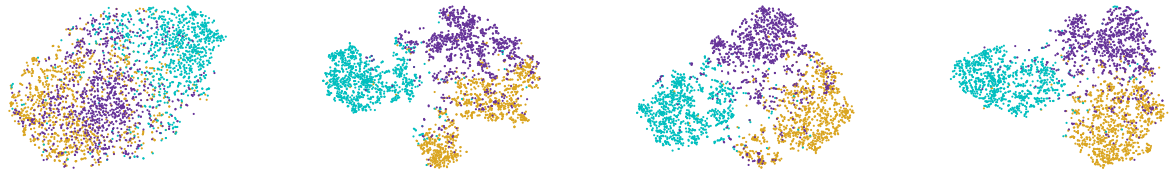


Figure 3. Classification results of our method at different parameter settings (*i.e.,* α, β, and λ) on all datasets.

D.1. Parameter Analysis

In the proposed method DMG, we employ the non-negative parameters (*i.e.,* α, β, and λ) to achieve a trade-off between each term of the objective function. To investigate the impact of α, β, and λ in Eq. (8) with different settings, we conduct the node classification on all multiplex graph datasets by varying the value of parameters in the range of $[10^{-3},10^{3}]$ and reporting the results in Figure 3. As shown in Figure 3, the proposed method DMG consistently achieves significant performance while the values of parameters are set appropriately (*e.g.,* $[10^{-2},10^{0}]$). Moreover, if the values of parameters are too large (*e.g.,* $> 10^{1}$) or too small (*e.g.,* $< 10^{-2}$), the proposed DMG obtains inferior performance due to the failure to obtain the disentangled common and private representations or the failure to preserve the complementarity and remove noise in the

private representations. This again validates that it is necessary to disentangle common and private representations, as well as to preserve complementarity and remove noise in private representations and validates the effectiveness of our method.



(a) w/o $\mathcal{L}_{cor}$ (SIL=0.145) (b) w/o $\mathcal{L}_{rec}$ (SIL=0.381) (c) w/o $\mathcal{L}_{con}$ (SIL=0.406) (d) with $\mathcal{J}$ (SIL=0.452)

Figure 4. Visualization plotted by t-SNE (and the corresponding silhouette scores (SIL) of node representations) for our method on the ACM dataset, where different colors represent different classes of the nodes. (a) DMG without $\mathcal{L}_{cor}$; (b) DMG without $\mathcal{L}_{rec}$; (c) DMG without $\mathcal{L}_{con}$; and (d) DMG with complete objective function.

D.2. Additional Ablation Study

D.2.1. Effectiveness of each component

In the main paper above, we investigate the effectiveness of each component of the objective function on the semi-supervised task (*i.e.,* node classification). To further verify and visualize the effectiveness of each component on unsupervised task, we investigate the performance of t-SNE visualization (Van der Maaten & Hinton, 2008) on different components of the objective function by reporting the results in Figure 4. Similar to the ablation study on the classification task, we have the observations as follows. First, our method with the complete objective function outperforms the variant without any loss, indicating that each component of the objective function is necessary for our method. Second, the variant without $\mathcal{L}_{cor}$ obtains the worst results, compared to the other two variants (without $\mathcal{L}_{rec}$ and without $\mathcal{L}_{con}$, respectively), indicating that the correlation loss may play an important role in our framework. The results on the t-SNE visualization are consistent with the ablation study of each component on the node classification in the main paper. Thus, the effectiveness of each component is further verified.

Table 8. Classification and clustering performance under different combinations of private representations of graphs.

Datasets	**Graphs**	**Node classification**		**Clustering**	
		Macro-F1	Micro-F1	Accuracy	NMI
ACM	PSP	90.3 ± 0.2	90.3 ± 0.4	92.1 ± 0.3	73.5 ± 0.4
	PAP	90.7 ± 0.4	90.6 ± 0.5	92.6 ± 0.3	74.1 ± 0.2
	PSP+PAP	$\mathbf{91.0 \pm 0.3}$	$\mathbf{90.9 \pm 0.4}$	$\mathbf{92.9 \pm 0.3}$	$\mathbf{74.5 \pm 0.4}$
IMDB	MAM	55.6 ± 0.4	57.5 ± 0.5	59.9 ± 0.3	16.3 ± 0.2
	MDM	57.1 ± 0.3	58.3 ± 0.3	60.2 ± 0.4	16.9 ± 0.2
	MAM+MDM	$\mathbf{57.6 \pm 0.2}$	$\mathbf{58.9 \pm 0.4}$	$\mathbf{60.3 \pm 0.5}$	$\mathbf{17.0 \pm 0.3}$

D.2.2. Effectiveness of private representations in each graph

With the complementarity and noise constraint in our framework, the private representations learned by our method are expected to contain the complementary contents of each graph. Therefore, we investigate the classification and clustering performance with private representations belonging to different graphs, and report the results in Table 8. From Table 8, we have observations as follows. First, private representations belonging to different graphs tend to have different importance. For example, our method only using the private representations of MDM graph outperforms our method only using the MAM graph on the ACM dataset on all downstream tasks. This makes sense as the complementary contents in a certain graph may be more than others. Second, our method with two graphs outperforms our method with only one graph on all downstream tasks. This verifies again that our method is available to preserve the complementarity in different graphs to benefit downstream tasks.

Table 9. Classification performance (*i.e.,* Macro-F1 and Micro-F1) on all multiplex graph datasets.

Method	ACM		IMDB		DBLP		Freebase	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
DMG-HR	76.6 ± 0.5	76.7 ± 0.4	43.1 ± 0.3	44.6 ± 0.4	88.8 ± 0.4	90.2 ± 0.5	48.1 ± 0.4	49.5 ± 0.5
DMG-NCE	88.2 ± 0.3	88.1 ± 0.4	55.2 ± 0.5	56.9 ± 0.3	92.7 ± 0.3	93.6 ± 0.2	59.4 ± 0.4	64.0 ± 0.3
DMG	$\mathbf{91.0 \pm 0.3}$	$\mathbf{90.9 \pm 0.4}$	$\mathbf{57.6 \pm 0.2}$	$\mathbf{58.9 \pm 0.4}$	$\mathbf{93.3 \pm 0.2}$	$\mathbf{94.0 \pm 0.3}$	$\mathbf{62.4 \pm 0.7}$	$\mathbf{65.9 \pm 0.8}$

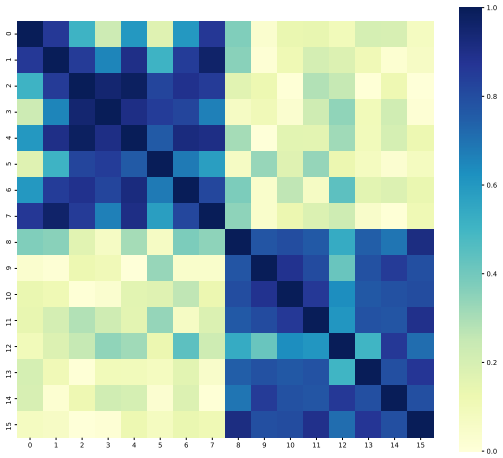


Figure 5. The correlation map of common and private representations learned by the proposed DMG on the ACM dataset.

D.3. Visualization of Disentangled Representation Learning

To verify the effectiveness of disentangled representation learning, *i.e.,* the common representations are clean to private representations, we follow the literature (Ma et al., 2019; Li et al., 2021) to calculate the correlation between the common representations and the private representations. Specifically, the smaller the correlation between the common representation and the private representation, the cleaner the common information learned by the representation learning method. To do this, we implement our method on the ACM dataset, and visualized the correlation of learned representations (concatenated by common and private representations, dimensions of them are both set as 8) in Figure 5. More specifically, the correlation map includes four blocks, where both the upper right block and the lower left block indicate the correlation between the common representations and the private representations. Based on Figure 1, their correlations are small, so the common information learned by our method is clean, and the effectiveness of disentangled representation learning is verified.

D.4. Comparison with Hard Matching Loss and InfoNCE Loss

In the proposed DMG, we investigate the matching loss to align the common representations from different graphs with a common variable $\mathbf{S}$. Actually, the matching loss in our framework can be regarded as a soft regularization to the common representations from different graphs, through a bridge (*i.e.,* common variable $\mathbf{S}$). Therefore, we consider replacing the soft matching with hard regularization (Chen et al., 2020) by removing the common variable, formulated as

$$\begin{aligned} &\mathcal{L}_{HR} = \frac{1}{N}\sum_{r=1}^{\mathcal{R}}\sum_{n=1}^{N}(\mathbf{c}_n^{(r)} - \mathbf{c}_n^{(r')})^2, \\ &\text{s.t. } \frac{1}{N}\sum_{n=1}^{N}\mathbf{c}_n\mathbf{c}_n^\top = \mathbf{I}, \end{aligned} \tag{31}$$

where $r \neq r'$. Moreover, in the proposed DMG, we investigate the contrastive loss to preserve the complementarity and remove noise in the private representations. Then we consider replacing it with another widely-used contrastive loss,

i.e., InfoNCE loss (Oord et al., 2018), formulated as

$$\mathcal{L}_{NCE} = \frac{1}{\mathcal{R}} \frac{1}{N} \sum_{r=1}^{\mathcal{R}} \sum_{i=1}^{N} \log \frac{e^{\theta\left(\boldsymbol{u}_i^{(r)}, \boldsymbol{v}_i^{(r)}\right)}}{\frac{1}{N} \sum_{j=1}^{N} e^{\theta\left(\boldsymbol{u}_i^{(r)}, \boldsymbol{v}_j^{(r)}\right)}}, \tag{32}$$

where $(\boldsymbol{u}_i^{(r)}, \boldsymbol{v}_i^{(r)})$ and $(\boldsymbol{u}_i^{(r)}, \boldsymbol{v}_j^{(r)})$ indicate positive and negative pairs, respectively. In our implementation, we replace the positive pair and negative pair with the node pairs in $\mathcal{E}_c^{(r)}$ and $\mathcal{E}_n^{(r)}$. Then we denote the modified models as DMG-HR and DMG-NCE, respectively, and report the classification performance of the modified models under identical settings with the original model DMG in Table 9.

From Table 9, we have the following observations. First, compared to the variant model DMG-HR, our method on average improves by 21.9%, on all multiplex graph datasets. The reason for this may be that using a hard regularization will align the common representations directly at the initial stage of training, but the common representations may not be optimal at this point. In addition, it is easier to enforce the orthogonality and zero-mean constraints on common variable than on common representations. Second, compared to the variant model DMG-NCE, our method also achieves superior performance on all multiplex graph datasets. The results empirically demonstrate that, although InfoNCE is a strict estimator of the mutual information, the contrastive loss in our framework is more effective and shows better downstream performance.

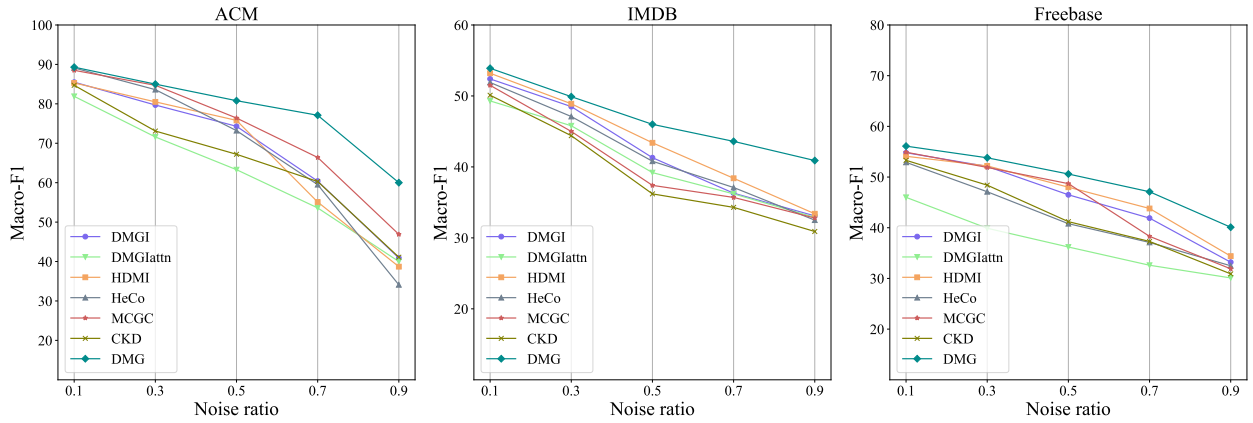


Figure 6. Classification performance of our method and all self-supervised MGRL methods under different noisy edges ratio η on ACM, IMDB, and Freebase dataset.

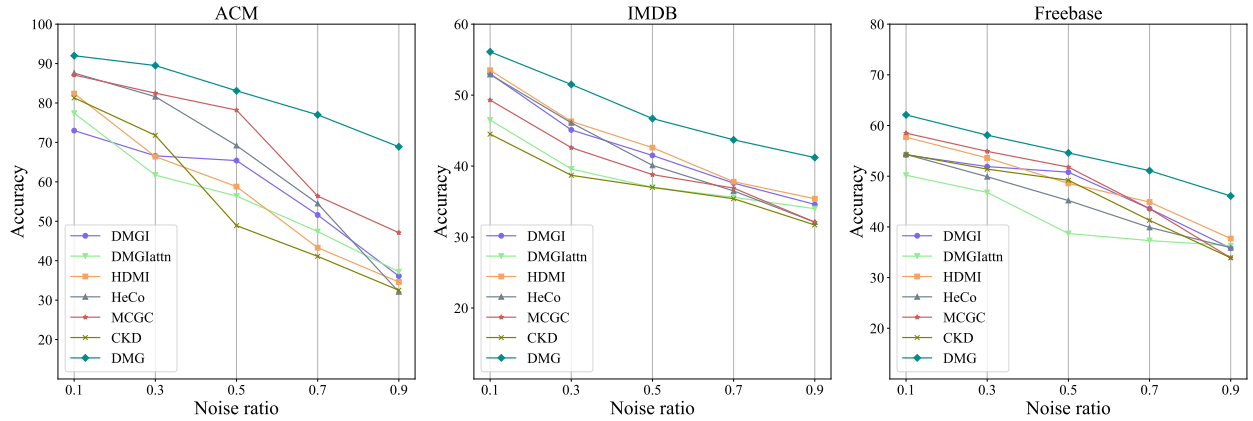


Figure 7. Clustering performance of our method and all self-supervised MGRL methods under different noisy edges ratio η on ACM, IMDB, and Freebase dataset.

E. Discussion on Potential Limitations and Future Directions

This section discusses the limitations of the proposed method and related future directions.

E.1. Features Multiplicity

Currently, our method mainly considers the private information in different graph structures since node features of different graphs are generated from a shared feature matrix. However, if node features in different graphs are also multiplexed, *i.e.,* the node features in each graph reveal different self-properties of the nodes, although our method can still extract complete and clean common information of different graphs, it cannot further explicitly preserve the complementarity and remove the noise in node features of different graphs. Therefore, future research may consider the multiplicity of node features so as to capture the private information in node features and to further preserve the complementarity and remove noise in node features of different graphs.

E.2. Unattributed Graphs

Currently, our method mainly considers graph datasets with node features. We should note that there exist cases where the nodes are unattributed and all information is contained in the graph topology, especially for some graph-level datasets. In fact, in our experimental section, we have chosen the unattributed graph dataset (*i.e.,* Freebase) to further verify the effectiveness and robustness of our method by simply assigning one-hot vectors as node features. Obviously, we can observe that our method obtains more significant improvements on the Freebase dataset than other multiplex graph datasets with node features. This may indicate that removing noise and preserving complementarity in graph structures is more important to unattributed graph datasets. Therefore, future research may consider designing methods specifically for unattributed graph datasets to further improve their effectiveness and robustness.