

Balanced Multi-Relational Graph Clustering

Zhixiang Shen

University of Electronic Science and
Technology of China
Chengdu, Sichuan, China
zhixiang.zxs@gmail.com

Haolan He

University of Electronic Science and
Technology of China
Chengdu, Sichuan, China
HaolanHe7777@gmail.com

Zhao Kang*

University of Electronic Science and
Technology of China
Chengdu, Sichuan, China
zkang@uestc.edu.cn

Abstract

Multi-relational graph clustering has demonstrated remarkable success in uncovering underlying patterns in complex networks. Representative methods manage to align different views motivated by advances in contrastive learning. Our empirical study finds the pervasive presence of imbalance in real-world graphs, which is in principle contradictory to the motivation of alignment. In this paper, we first propose a novel metric, the Aggregation Class Distance, to empirically quantify structural disparities among different graphs. To address the challenge of view imbalance, we propose Balanced Multi-Relational Graph Clustering (BMGC), comprising unsupervised dominant view mining and dual signals guided representation learning. It dynamically mines the dominant view throughout the training process, synergistically improving clustering performance with representation learning. Theoretical analysis ensures the effectiveness of dominant view mining. Extensive experiments and in-depth analysis on real-world and synthetic datasets showcase that BMGC achieves state-of-the-art performance, underscoring its superiority in addressing the view imbalance inherent in multi-relational graphs. The source code and datasets are available at <https://github.com/zxlearningdeep/BMGC>.

CCS Concepts

• **Computing methodologies** → **Cluster analysis**; *Neural networks*; *Regularization*.

Keywords

Multi-view Graph Clustering, Imbalanced Graph Learning, Graph Representation Learning, Graph Homophily

ACM Reference Format:

Zhixiang Shen, Haolan He, and Zhao Kang. 2024. Balanced Multi-Relational Graph Clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, October 28–November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3664647.3681325>

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0686-8/24/10

<https://doi.org/10.1145/3664647.3681325>

1 Introduction

Multi-relational graphs, which involve a set of nodes with multiple relations, are prevalent in the real world because of their extraordinary ability in characterizing complex systems [29]. Some typical instances are citation networks, social networks, and knowledge graphs [24, 38]. Recently, the unsupervised exploration of the inherent pattern in complex networks has attracted considerable attention, particularly in the context of multiview graph clustering (MVGC). Conventional MVGC techniques typically combine graph optimization with clustering techniques such as subspace clustering and spectral clustering [13, 22]. With the advancement of Graph Neural Networks (GNNs), a new wave of deep MVGC methods has been proposed, such as O2MAC [5], DMGI [25], HDMI [11], BTGF [28]. They have demonstrated significant efficacy.

However, representative MVGC methods often align all views to seek consistent information with the aid of a contrastive learning mechanism [11, 20, 28, 36]. This approach often ignores the fact that different views in real-world data do not always carry equal significance, i.e., the imbalance phenomenon. Our empirical analysis of real-world multi-relational graphs confirms this intuition. As shown in Fig. 1, different relations exhibit a big gap in classification accuracy. Therefore, naively aligning different views could degrade the final performance.

To this end, we address the view imbalance problem in multi-relational graphs in this work. Unlike other multiview data, the view differences in multi-relational graphs are rooted in their topology structures. Thus, a natural question arises: (Q1) **How can we quantify the structural disparities between views in multi-relational graphs?** Previous studies in multimodal learning indicate the presence of a dominant view in view-imbalanced data [34]. Given the clustering tasks, another question appears: (Q2) **How can we discover the dominant view without supervision to guide multi-relational graph clustering?**

In addressing Q1, we propose a simple yet effective view evaluation metric: Aggregation Class Distance (ACD). Unlike previous methods that solely calculate graph homophily ratios at the edge or node level [23], ACD takes into account both the aggregation process and the distribution of node classes. Empirical studies conducted on real-world datasets demonstrate the efficacy of this novel metric in evaluating the quality of views in multi-relational graphs.

For Q2, we propose Balanced Multi-Relational Graph Clustering (BMGC), which incorporates unsupervised dominant view mining and dual signal guided representation learning. A dynamic method of unsupervised exploration of the dominant view is employed throughout the training process, taking advantage of view-specific representations and original node features. Theoretical analysis establishes the connection between this approach and ACD, ensuring the effectiveness of dominant view mining. Afterward, dual signals

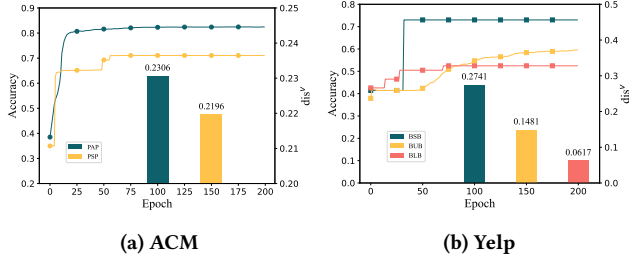


Figure 1: The node classification accuracy for each view of the test set, along with the corresponding ACD. The trends in accuracy and ACD are consistent.

from the dominant view and node features supervise the graph embedding, promoting co-aligned representation learning. Finally, the dominant assignment is utilized to further enhance the clustering performance. In particular, dynamic extraction of the dominant view coupled with representation learning synergistically boosts model training.

We underscore the primary contributions of this study as follows:

- We explore view imbalance in multi-relational graphs and design a metric for evaluating view quality. Our empirical study confirms the presence of view disparities and validates the utility of our proposed metric.
- To our best knowledge, we are the first to tackle view imbalance in multi-relational graphs without supervision. BMGC integrates an unsupervised method for dominant view mining and dual signal guided representation learning. Furthermore, the dominant assignment is leveraged to facilitate self-training clustering. A theoretical guarantee is provided to demonstrate the efficacy of our approach.
- We conduct comprehensive experiments and in-depth analysis on real-world and synthetic datasets. BMGC surpasses existing advanced methods across all datasets, affirming its effectiveness and superiority in addressing view imbalance.

2 Related Work

2.1 Multiview Graph Clustering

Recently, there have been extensive explorations into multiview graph clustering. Typical shallow methods, such as MvAGC [13] and MCGC [22], combine graph filtering with self-expression learning to leverage attribute and structural information simultaneously.

With the progress in representation learning, several deep methods have emerged. Most of them start by unsupervised learning of node representations and then apply the k-means algorithm on these representations to obtain clustering results. O2MAC [5] is the first to employ GNN for MVGC, selecting the most informative view as input and reconstructing the graph structures of all views to capture shared information. Although O2MAC considers discrepancies in different graph structures, its encoding strategy, which retains only the best view, results in degradation into a single-view method. Moreover, it uniformly reconstructs the graph structures of all views, which in turn disregards the view imbalance, further leading to suboptimal results. DMGI [25] and HDMI [11] optimize

embeddings by maximizing mutual information between local and global representations. MGCCN [15], MGDCR [20], and BTGF [28] incorporate various contrastive losses to achieve the alignment of the representation and prevent dimension collapse. DualLGR [14] extracts supervised signals from node attributes and graph structures to guide the MVGC. CoCoMG [26] and DMG [21] approach multi-view representation learning from the perspectives of consistency and complementarity. Although numerous methods achieve representation learning through multiview alignment, most of them overlook the inherent performance disparities between different views. These alignment-based methods tend to treat all graphs equally, which compromises the quality of node representations and thereby deteriorates the clustering results.

In supervised or semi-supervised tasks, methods like HAN [35] and SSAMN [30] consider the varying importance of views. However, they require labeled information for training, which is unsuitable for unsupervised tasks. **To our best knowledge, we are the first to address view imbalance in multiview graph clustering.**

2.2 Imbalanced Multiview Learning

Numerous efforts have been dedicated to addressing the challenges of imbalanced multiview learning from diverse perspectives. Works such as [8, 16] tackle imbalanced views through decision-level fusion. Specifically, they initially cluster each view and then fuse the view-specific clustering results. Another distinct avenue involves leveraging similarity graphs. MDcR [40] constructs balanced view-specific inter-instance similarity graphs, utilizes embedding techniques to acquire latent representations, and concatenates them to form the final representation for clustering. In contrast, FMUGE [39] takes a different approach to model order, initially combining view-specific similarity matrix to create a common similarity graph, followed by learning a comprehensive multiview representation. However, all of these methods cannot handle graph structure data.

Imbalanced multimodal learning has also attracted widespread attention. Recent theoretical advancements have demonstrated the potential of multimodal learning to surpass the upper limits of single-modal performance [10]. However, due to the varying confidence levels and noise across different modalities, the learning process is susceptible to inducing bias towards a dominant modality. To achieve a balanced multimodal classification, OGM [27] devises a modality-wise difference ratio to monitor the contribution discrepancy of each modality to the target, thus adaptively adjusting the gradients of each modality. Subsequently, PMR [6] proposes Prototypical Modality Rebalance, accelerating the slow-learning modality by enhancing its clustering towards prototypes. Despite the effectiveness of these methods, none have considered graph data. Therefore, methods to handle the view imbalance in the realm of unsupervised multiview graph learning are urgently needed.

3 Empirical Study

Notation. In this work, we define a multi-relational graph as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}_1, \dots, \mathcal{E}_v, \dots, \mathcal{E}_V, X\}$, where \mathcal{V} is the node set with N nodes and \mathcal{E}_v is the edge set in the v -th view. $V > 1$ is the number of relational graphs. $X \in \mathbb{R}^{N \times d_f}$ is the feature matrix and $x \in \mathbb{R}^N$ is a column of the feature matrix that represents a graph signal. \tilde{A}^v denotes the original adjacency matrix of the v -th view.

D^v represents the degree matrix. The normalized adjacency matrix of the v -th view is given by $A^v = (D^v)^{-\frac{1}{2}} \tilde{A}^v (D^v)^{-\frac{1}{2}}$. It is a well-known fact that the eigenvalues of A^v in each view are contained within $[-1, 1]$. $\hat{A}^v = (D^v + I)^{-\frac{1}{2}} (\tilde{A}^v + I) (D^v + I)^{-\frac{1}{2}}$ represents the normalized adjacency matrix with a self-loop to each node, where I is an identity matrix. C is the number of node classes, and $y \in \mathbb{R}^N$ denotes the label vector.

In a multi-relational graph, the imbalance between views stems from differences in graph structure: some graphs contain more task-relevant information, while others are less task-relevant. Previous research has analyzed the impact of the graph structure on GNN from the perspective of graph homophily, suggesting that structures with high homophily ratios often exhibit superior performance [3, 42]. Here, the edge-level homophily ratio (hr) is defined as $hr = \frac{1}{|\mathcal{E}|} \sum_{(i,j) \in \mathcal{E}} \mathbb{1}(y_i = y_j)$, where $\mathbb{1}(\cdot)$ is the indicator function that equals 1 if its argument is true and 0 otherwise. However, recent studies have shown that neighbors of different classes may not necessarily make the nodes indistinguishable [19]. Graph structure analysis should consider node neighborhood patterns and the aggregation process. To quantify structural disparities across different views, we propose a simple yet effective metric: *Aggregation Class Distance* (ACD). ACD evaluates structure quality based on aggregated feature distribution of node classes, adapting better to real-world complexity than assuming a direct correlation between homophily ratio and task performance. The theoretical analysis in Section 5 substantiates this assertion. We choose the Simple Graph Convolution (SGC) as the aggregation method [37], a widely used representative aggregation operation [3, 19]. The ACD is defined as follows.

DEFINITION 1. *The aggregation class distance for the v -th view, denoted as dis^v , is calculated as:*

$$X^v = (A^v)^K X, \quad \bar{X}_m^v = \frac{1}{N_m} \sum_{y_i=m} X_i^v \quad (1)$$

$$dis^v = \frac{2}{C^2 - C} \sum_{m=1}^C \sum_{n=m+1}^C \|\bar{X}_m^v - \bar{X}_n^v\| \quad (2)$$

where N_m is the number of nodes in class m and K denotes the radius of aggregation. The reciprocal of $\frac{2}{C^2 - C}$ represents the computation count for pairwise inter-class distances.

\bar{X}_m^v represents the centroid of aggregated features for nodes with class m in the v -th view. The metric dis^v gauges the inter-class distance of aggregated features. A higher value indicates better discriminability among different classes.

To demonstrate the connection between ACD and view performance, we conduct empirical research on real-world datasets. We randomly select 30% of the nodes as the training set, leaving the remaining ones for the test set. The aggregation radius is set to 3, aligning with the common layer count of many GNN models [2]. A linear layer serves as the classifier. As shown in Fig. 1, the line represents the classification accuracy of each view, while the bar chart indicates the corresponding ACD. Different views yield different results, affirming the existence of view imbalance. In each dataset, the performance of one view significantly exceeds that of others, and we refer to it as the dominant view. Furthermore, views with higher ACD values exhibit better classification results,

underscoring the efficacy of ACD in gauging structural disparities between views. This empirical evidence supports the notion that ACD serves as a valuable metric for evaluating the quality of views in multi-relational graphs. It is worth noting that ACD uses node labels to assess the graph structure quality of each view, meaning that it is a supervised approach and cannot be directly applied to unsupervised learning tasks like clustering.

4 Methodology

In this section, we propose Balanced Multi-Relational Graph Clustering, as depicted in Fig. 2, to overcome inherent view imbalance.

4.1 Scalable Graph Encoding

Unlike most GNN-based approaches [20, 21], we decouple graph propagation and dimensionality reduction to improve scalability. Initially, we perform propagation on the node features separately for each view, acquiring view-specific aggregated features. Similarly to the approximate personalized propagation in [4], we introduce the features of the original node as a teleport vector in each layer of the propagation process:

$$X^{v,0} = X, \quad X^{v,k+1} = (1 - \alpha) \hat{A}^v X^{v,k} + \alpha X \quad (3)$$

where X acts as both the starting matrix and the teleport set for each view. The hyper-parameter $\alpha \in [0, 1]$ represents the teleport probability. $k \in [0, K - 1]$ and the aggregated features $X^v = X^{v,K}$. These features are then fed into a shared encoder for dimensionality reduction:

$$Z^v = f_\Theta(X^v) \quad (4)$$

where $Z^v \in \mathbb{R}^{N \times d_r}$ denotes node representations in the v -th view. This decoupled setup avoids the time-consuming graph convolution operations during training. Subsequently, the representations for each view are fed into a shared decoder for the reconstruction of view-specific aggregated features. Effective training of each view is ensured by optimizing the reconstruction cosine error:

$$\hat{X}^v = g_\Theta(Z^v) \quad (5)$$

$$\mathcal{L}_{REC} = \frac{1}{NV} \sum_{v=1}^V \sum_{i=1}^N \left(1 - \frac{(X_i^v)^\top \hat{X}_i^v}{\|X_i^v\| \cdot \|\hat{X}_i^v\|} \right) \quad (6)$$

where the encoder $f_\Theta(\cdot)$ and the decoder $g_\Theta(\cdot)$ are both implemented using a Multilayer Perceptron (MLP) in this study.

4.2 Unsupervised Dominant View Mining

Due to the absence of label information, we cannot directly utilize ACD for the assessment of view quality in multi-relational graph clustering. Therefore, grounded in the principle of invariance in the distribution of similarity between instances, we propose an unsupervised method for dominant view mining:

$$v^* = \arg \min_v \|XX^\top - X^v(X^v)^\top\|_F^2 \quad (7)$$

where v^* denotes the dominant view. The above expression quantifies the discrepancy between the similarity matrices of original features and view-specific aggregated features, henceforth referred to as the “unsupervised metric”. Therefore, the dominant view

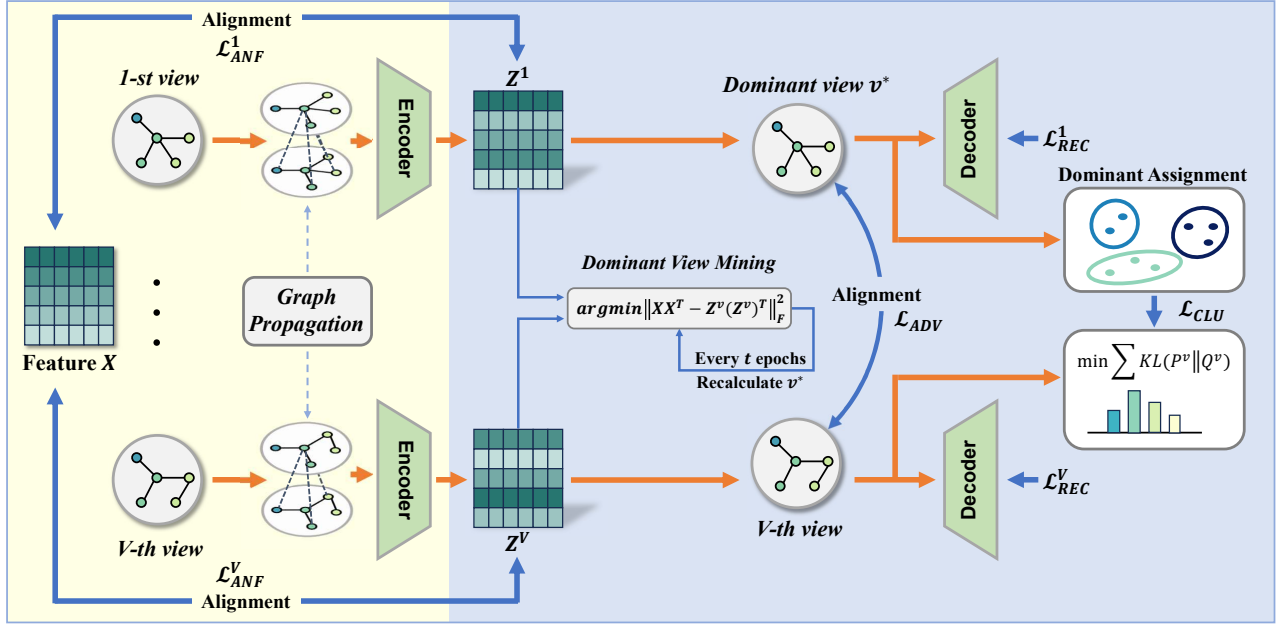


Figure 2: Illustration of our proposed framework BMGC. Firstly, node representations for each view are obtained through scalable graph encoding. Then, unsupervised dominant view mining and dual signals guided representation learning synergistically facilitate model training. Finally, the dominant assignment further enhances clustering quality.

should optimally maintain instance similarities. In Section 5, we theoretically establish the effectiveness of our approach.

Considering potential noise in real-world data, we refrain from directly using aggregated features to gauge view quality and, instead, rely on node representations:

$$v^* = \arg \min_v \|XX^T - Z^v(Z^v)^T\|_F^2 \quad (8)$$

In the training process, we initialize the dominant view using Equation (7) and recalculate it every t epochs using Equation (8). As training progresses, the quality of node representations improves, thereby bolstering the reliability of dominant view mining. Simultaneously, the dominant view would guide representation learning, as elaborated later. It constitutes a mutually reinforcing process.

4.3 Co-aligned Representation Learning

After determining the dominant view, we use it to improve the representation quality. We employ contrastive learning to align the representations of other views with the dominant view. The representations of each view are projected to a shared latent space using separate learnable MLPs for fair similarity measurement and loss calculation. The contrastive loss is defined as follows:

$$\ell(Z_i^v, Z_i^{v^*}) = -\log \frac{e^{\text{sim}(\tilde{Z}_i^v, \tilde{Z}_i^{v^*})/\tau}}{\sum_{j=1}^N e^{\text{sim}(\tilde{Z}_i^v, \tilde{Z}_j^{v^*})/\tau}} \quad (9)$$

where \tilde{Z}_i^v is the non-linear projection of Z_i^v . $\text{sim}(\cdot)$ refers to the cosine similarity and τ is the temperature parameter. The loss for aligning with the dominant view is given by:

$$\mathcal{L}_{ADV} = \frac{1}{2N(V-1)} \sum_{v=1}^V \sum_{v \neq v^*}^N (\ell(Z_i^v, Z_i^{v^*}) + \ell(Z_i^{v^*}, Z_i^v)) \quad (10)$$

Each view, along with the supervision from the dominant view, should also preserve a consistent similarity distribution among the nodes. Hence, we introduce a loss to ensure alignment with the node features:

$$\mathcal{L}_{ANF} = \frac{1}{N^2V} \sum_{v=1}^V \|XX^T - Z^v(Z^v)^T\|_F^2 \quad (11)$$

Note that the loss \mathcal{L}_{ANF} shares a similar form with the unsupervised metric proposed in Section 4.2. This establishes a foundation for ensuring the reliability of our approach in continuously mining the dominant view during training. Ultimately, guided by dual signals from both the dominant view and node features, we accomplish the co-aligned representation learning:

$$\mathcal{L}_{CAL} = \mathcal{L}_{ADV} + \mathcal{L}_{ANF} \quad (12)$$

4.4 Dominant Assignment Enhanced Clustering

Most deep clustering methods leverage target distribution and soft cluster assignment probability distributions to achieve a self-training clustering scheme, with the cluster distribution typically obtained by applying k-means [17, 28, 31]. To improve the clustering performance, we substitute representation distributions in other views with cluster assignments derived from the dominant

view. Specifically, we apply k-means to the representations of the dominant view to obtain the dominant assignment:

$$\hat{y} = KMeans(\{Z_i^v : i = 1, \dots, N\}) \quad (13)$$

Then, the soft assignment distribution Q^v in the v -th view can be formulated as:

$$\sigma_j^v = \frac{1}{N_j} \sum_{\hat{y}_i=j} Z_i^v, \quad q_{ij}^v = \frac{(1 + \|Z_i^v - \sigma_j^v\|^2)^{-1}}{\sum_{k=1}^C (1 + \|Z_i^v - \sigma_k^v\|^2)^{-1}} \quad (14)$$

where N_j denotes the number of nodes with the cluster assignment j and q_{ij}^v is measured using Student's t-distribution to denote the similarity between representation Z_i^v and the clustering center σ_j^v . The target distribution P^v is computed as:

$$p_{ij}^v = \frac{(q_{ij}^v)^2 / \sum_{i=1}^N q_{ij}^v}{\sum_{k=1}^C \left((q_{ik}^v)^2 / \sum_{i=1}^N q_{ik}^v \right)} \quad (15)$$

We minimize the KL divergence between the distributions Q^v and P^v for each view to enhance cluster cohesion. The final node representation $Z = [Z^1, \dots, Z^V] \in \mathbb{R}^{N \times Vd_r}$ is obtained by concatenating representations from all views. We simultaneously minimize the KL divergence between the Q and P distributions of the final representation Z :

$$\mathcal{L}_{CLU} = KL(P||Q) + \frac{1}{V} \sum_{v=1}^V KL(P^v||Q^v) \quad (16)$$

The overall objective of BMGC, which we aim to minimize through the gradient descent algorithm, consists of three loss terms:

$$\mathcal{L} = \mathcal{L}_{REC} + \mathcal{L}_{CAL} + \mathcal{L}_{CLU} \quad (17)$$

For large-scale datasets, our method, which benefits from scalable graph encoding, eliminates the need for neighbor sampling during the training process. Consequently, we can directly perform mini-batch training, where all loss terms are computed solely from nodes within the batch.

5 Theoretical Analysis

In this section, we use a synthetic network to theoretically demonstrate the efficacy of BMGC in extracting the dominant view. For simplicity, we adopt SGC as the feature aggregation method.

Data Assumption. A multi-relational graph G has N nodes partitioned into 2 equally sized communities C_1 and C_2 . Let $c_1, c_2 \in \{0, 1\}^N$ be indicator vectors for membership in each community, that is, the j^{th} entry of c_i is 1 if the j^{th} node is in C_i and 0 otherwise. G has V views, each is generated by SBM [1], with intra- and inter-community edge probabilities p^v and q^v . G is such a graph model with a feature matrix $X = F + H \in \mathbb{R}^{N \times d_f}$, where each column of H follows a zero-centered, isotropic Gaussian noise distribution $\mathcal{N}(0, \sigma^2 I)$ and these columns are mutually independent. The matrix F is defined as $F = c_1 \mu_1^T + c_2 \mu_2^T$, where $\mu_1, \mu_2 \in \mathbb{R}^N$ has the same Euclidean norm $\|\mu\|$, representing the expected characteristic vector of each community. In addition, let $\bar{\mu} = \frac{1}{2}(\mu_1 + \mu_2)$ be the average of the feature vector means.

LEMMA 1. Let X^v be the aggregated feature matrix of the v -th view by applying SGC, with the number of hops K , to the expected adjacency matrix \tilde{A}^v and the feature matrix X . Then, $X^v = F^v + c_1(\theta_1^v)^T + c_2(\theta_2^v)^T$, where $F^v = (\lambda_2^v)^K F + (1 - (\lambda_2^v)^K)(\bar{\mu}^T)$, θ_1^v and $\theta_2^v \in \mathbb{R}^{d_f}$ are both distributed according to $\mathcal{N}(0, \frac{1}{N}(1 + (\lambda_2^v)^{2K})\sigma^2 I)$, and $\lambda_2^v = \frac{p^v - q^v}{p^v + q^v} \in [-1, 1]$ is the second largest non-zero eigenvalue of the associated normalized adjacency matrix A^v .

THEOREM 1. Let \bar{X}_1^v and \bar{X}_2^v denote the centroid of aggregated features for each community in the v -th view. Then, $\mathbb{E}[\bar{X}_1^v - \bar{X}_2^v] = (\lambda_2^v)^K(\mu_1 - \mu_2)$ and $\mathbb{E}[XX^T - X^v(X^v)^T] = \frac{1 - (\lambda_2^v)^{2K}}{2}(\|\mu\|^2 - \mu_1^T \mu_2)(c_1 c_1^T + c_2 c_2^T - c_1 c_2^T - c_2 c_1^T) + \omega(\sigma^2)$, where $\omega(\sigma^2)$ represents the sum of terms containing σ^2 that are of negligible magnitude.

When negligible terms are ignored, it becomes clear that the unsupervised identification of the dominant view essentially corresponds to the view with the maximum $(\lambda_2^v)^2$. Additionally, under this data assumption, $dis^v = \|\bar{X}_1^v - \bar{X}_2^v\|$ indicates that a view with a larger $(\lambda_2^v)^2$ is more likely to exhibit a larger ACD. Specifically, the consistent changes in both dis^v and $\|XX^T - X^v(X^v)^T\|_F^2$ related to $(\lambda_2^v)^2$ reveal that the identified dominant view is the one with the largest ACD value.

The theoretical findings are interpretable. λ_2^v is influenced by the ratio of intra- to inter-community edge probabilities, essentially measuring the homophily level of the graph structure in the v -th view. It tends toward 1 for complete homophily and -1 for complete heterophily. In both cases where $(\lambda_2^v)^2$ approaches 1, there would be no confusion between nodes from different communities. For example, when $\lambda_2^v = -1$, if K is odd, feature exchange occurs between the two communities after aggregation; if K is even, the features would remain unchanged. In such cases with pure graph structures, the view consistently demonstrates a larger ACD and a smaller unsupervised metric, while the homophily ratio would tend toward zero as λ_2^v approaches -1 . On the contrary, when λ_2^v nears 0, the graph structure becomes uninformative, resulting in the view with a smaller ACD and a larger unsupervised metric. In summary, we draw two conclusions: **1. ACD is more universally applicable than the traditional homophily ratio in assessing the relevance between graph structures and downstream tasks. 2. The dominant view, mined through our unsupervised method, essentially corresponds to the view with the maximum ACD value, which ensures the effectiveness of unsupervised dominant view mining.**

6 Experiments

6.1 Datasets and Metrics

Datasets. We employ five publicly available real-world benchmark datasets and a large-scale dataset. ACM [5], ACM2 [7], and DBLP [41] are citation networks. Yelp [18] and Amazon [9] are review networks. MAG [33] is a large-scale citation network, constituting the largest dataset in multi-relation graph clustering thus far.

Metrics. We adopt four popular clustering metrics, including Accuracy (ACC), Normalized Mutual Information (NMI), F1 score, and Adjusted Rand Index (ARI). A higher value of them indicates a better performance.

Table 1: Clustering results on real-world datasets. The best and second-place results are highlighted using bold and underline, respectively. The asterisk (*) denotes the supervised baseline.

Datasets	Metric	HAN* 2019	VGAE 2016	DGI 2019	O2MAC 2020	DMGI 2020	MvAGC 2021	HDMI 2021	MCGC 2021	MGDCR 2023	DuaLGR 2023	DMG 2023	BTGF 2024	BMGC
Amazon	NMI	<u>0.4037</u>	0.0163	0.0532	0.1344	0.2623	0.2322	0.3702	0.2149	0.0318	0.2767	0.1218	0.3853	0.5768
	ARI	<u>0.4241</u>	0.0129	0.0202	0.0898	0.2605	0.1141	0.2735	0.1056	0.0055	0.2715	0.0283	0.2829	0.5626
	ACC	<u>0.7437</u>	0.3194	0.3762	0.4428	0.5581	0.5188	0.5251	0.4683	0.3489	0.6123	0.3887	0.6603	0.7856
	F1	<u>0.7433</u>	0.2725	0.2859	0.4424	0.5463	0.5072	0.5448	0.4804	0.2039	0.6215	0.3441	0.6612	0.7851
ACM	NMI	0.6864	0.491	0.6364	0.6923	0.6441	0.6735	0.645	0.7126	0.721	0.7328	0.7561	<u>0.758</u>	0.7841
	ARI	0.7489	0.5448	0.6822	0.7394	0.6729	0.7212	0.674	0.7627	0.6496	0.7942	0.8033	<u>0.8085</u>	0.8329
	ACC	0.9088	0.8228	0.8816	0.9042	0.8724	0.8975	0.874	0.9147	0.919	0.9271	0.9302	<u>0.9322</u>	0.9413
	F1	0.9085	0.8231	0.8829	0.9053	0.8709	0.8986	0.872	0.9155	0.8678	0.927	0.9306	<u>0.9331</u>	0.9416
DBLP	NMI	0.6998	0.6934	0.6168	0.7294	0.7489	0.7723	0.6361	0.6561	0.7595	0.7559	<u>0.7907</u>	0.6027	0.8013
	ARI	0.7641	0.7413	0.5653	0.7783	0.8032	0.828	0.6145	0.7088	0.8072	0.8168	<u>0.8384</u>	0.6534	0.8539
	ACC	0.9015	0.8868	0.7446	0.9071	0.9159	0.9284	0.7832	0.8752	0.9182	0.9242	<u>0.9344</u>	0.8509	0.9401
	F1	0.8966	0.8748	0.7392	0.901	0.9075	0.9231	0.7372	0.8186	0.9123	0.918	<u>0.9303</u>	0.8456	0.9364
ACM2	NMI	0.6435	0.4507	0.5779	0.4223	0.574	0.1819	0.5902	0.5307	0.5447	0.5988	0.6341	<u>0.6483</u>	0.7285
	ARI	<u>0.6979</u>	0.4347	0.5174	0.4451	0.5243	0.1879	0.5472	0.4396	0.4372	0.6399	0.6726	<u>0.6776</u>	0.7601
	ACC	<u>0.8943</u>	0.7358	0.8114	0.7537	0.8148	0.5949	0.8258	0.7129	0.6838	0.8676	0.8796	0.8853	0.9185
	F1	<u>0.8955</u>	0.7101	0.8261	0.7418	0.8267	0.4484	0.8386	0.5809	0.5854	0.8653	0.8773	0.8887	0.9215
Yelp	NMI	<u>0.6762</u>	0.3919	0.3942	0.3902	0.3729	0.2439	0.3912	0.3835	0.4423	0.6621	0.391	0.4135	0.7173
	ARI	<u>0.7205</u>	0.4257	0.4262	0.4253	0.3418	0.2925	0.3922	0.3517	0.4647	0.6847	0.4261	0.3564	0.7381
	ACC	<u>0.9082</u>	0.6507	0.6529	0.6507	0.5893	0.6314	0.6452	0.6561	0.7271	0.8948	0.6512	0.7192	0.9151
	F1	<u>0.9163</u>	0.5674	0.5679	0.5674	0.4878	0.567	0.5874	0.5749	0.5443	0.9051	0.5679	0.7307	0.9246

6.2 Experimental Setup

Baselines. We compare BMGC with various baselines, including the supervised multiview graph method HAN [35], single-view graph clustering methods VGAE [12] and DGI [32], and multiview graph clustering methods O2MAC [5], DMGI [25], MvAGC [13], HDMI [11], MCGC [22], MGDCR [20], DuaLGR [14], DMG [21], and BTGF [28]. All other methods are unsupervised excluding HAN, which serves as the supervised baseline.

Parameter Setting. Our model is trained for 400 epochs using the Adam optimizer with a learning rate of $1e-2$. The weight decay of the optimizer is set to $1e-4$. The recalculation interval t for the dominant view is every 50 epochs. We set the representation dimension d_r to 64 for ACM2 dataset and 10 for the other datasets. The temperature parameter τ is fixed at 1. The radius of graph propagation, K , is fixed at 3 and the teleport probability α is tuned in $[0, 0.3, 0.5]$. All experiments are implemented on the PyTorch platform using an Intel(R) Xeon(R) Platinum 8352V CPU and a GeForce RTX 4090 24G GPU.

6.3 Evaluation on Real-world Datasets

To evaluate the performance of our model, we compare BMGC with multiple baselines on five real-world datasets. For the supervised baseline HAN, we employ k-means on the node embeddings of the test set to yield clustering results. We conduct single-view clustering methods separately for each view and present the best results. Generally, BMGC consistently outperforms all compared methods regarding four metrics over all datasets. From Table 1, we have the following observations:

- The advantages of BMGC become evident when compared to other methods. In particular, our model significantly outperforms existing methods, including the supervised baseline, on Amazon and ACM2 datasets. Regarding second-place results on Amazon, our model improves NMI and ARI by 42.9% and 32.7%, respectively.
- In general, multiview graph methods outperform single-view methods like VGAE and DGI, demonstrating the superiority of multiview methods in graph clustering. However, in the Yelp dataset, most multiview baselines underperform compared to single-view methods, which may be attributed to the fact that these multiview methods overlook the imbalance among different views, leading to worse performance. Moreover, while the supervised baseline HAN surpasses the unsupervised baselines on most datasets, BMGC still outperforms it, underscoring the superiority of our method.
- Our model outperforms O2MAC which considers information differences among views. O2MAC retains only the most informative view while discarding others, which to some extent degenerates into a single-view method with worse results. Our model uses all the views and achieves better results by aligning the other views with the dominant view.

6.4 Evaluation on Synthetic Datasets

To further compare BMGC with other methods in addressing the imbalanced problem, we introduce a new synthetic dataset based on cSBM [4], named multi-relational cSBM. The multi-relational cSBM initially generates three views, each possessing unique graph structures with uniform homophily ratios, and sharing a common feature

Table 2: Clustering results on synthetic datasets.

Perturbation ratios	Metric	SGC			DMGI	HDMI	MGDCR	DuaLGR	DMG	BTGF	BMGC
		view 1	view 2	view 3	2020	2021	2023	2023	2023	2024	
20%	NMI	0.6142	0.5191	0.4973	0.675	0.5869	<u>0.8702</u>	0.4065	0.6326	0.7874	0.9209
	ARI	0.7207	0.6278	0.6053	0.7765	0.694	<u>0.9293</u>	0.5074	0.7321	0.8697	0.9612
	ACC	0.9245	0.8962	0.8891	0.9406	0.9165	<u>0.982</u>	0.8562	0.9278	0.9663	0.9902
50%	NMI	0.6142	0.3956	0.3896	0.6425	0.4766	<u>0.7959</u>	0.3314	0.5748	0.7213	0.8913
	ARI	0.7207	0.4959	0.4888	0.7471	0.583	<u>0.8761</u>	0.4229	0.683	0.8162	0.9432
	ACC	0.9245	0.8521	0.8496	0.9322	0.8818	<u>0.968</u>	0.8252	0.9132	0.9517	0.9856
100%	NMI	0.6142	0.2514	0.267	0.573	0.2821	<u>0.6887</u>	0.322	0.5195	0.6505	0.8178
	ARI	0.7207	0.327	0.3457	0.6812	0.3646	<u>0.7885</u>	0.4115	0.6257	0.7545	0.8926
	ACC	0.9245	0.7859	0.7941	0.9127	0.8019	<u>0.944</u>	0.8208	0.8955	0.9343	0.9724
150%	NMI	<u>0.6142</u>	0.1679	0.1564	0.4372	0.0374	0.5711	0.3079	0.4659	0.6104	0.7821
	ARI	<u>0.7207</u>	0.223	0.2086	0.5362	0.0441	0.6789	0.3953	0.5657	0.7175	0.8656
	ACC	<u>0.9245</u>	0.7361	0.7284	0.8662	0.5948	0.9121	0.8144	0.8761	0.9235	0.9652

matrix. All nodes are categorized into two classes. We randomly add noisy edges to two of these graphs to induce perturbations, where the perturbation ratio ρ controls the proportion of randomly added edges, simulating the imbalanced multi-relational graph. The undisturbed view is denoted as view 1, representing the dominant view, while the other two views are referred to as view 2 and view 3. Experiments are carried out for four ρ values: [20%, 50%, 100%, 150%].

To reveal the performance discrepancies of different views, we use SGC on each view to obtain view-specific aggregated features and then obtain clustering results through k-means. We select several representation learning-based approaches for comparison. The results, as shown in Table 2, indicate that a higher perturbation ratio leads to poorer performance for all methods. Our detailed observations are as follows.

First, the majority of multiview graph clustering methods yield unsatisfactory results. As the perturbation ratio increases, their performance degrades to a lower level. Second, as the perturbation ratio reaches 150%, the performance of the comparative methods even drops below the SGC result in view 1, indicating that when there is extensive noise in certain views of the dataset, the performance of multiview methods may deteriorate compared to the single view methods. In our method, aligning with the dominant view prevents the result from being impaired by low-quality views with noise. Third, our model maintains relatively stable performance as the perturbation ratio increases, with a maximal variation range of 17.7%, 11%, and 2.6% for NMI, ARI, and ACC respectively, showcasing the robustness for noisy data.

6.5 Evaluation on Large-scale Dataset

To evaluate the efficiency of BMGC, we conduct experiments on a large-scale multi-relational graph MAG. We select some scalable representation learning-based methods as baselines, while the remaining models run out of memory. We set the representation

Table 3: Quantitative results with standard deviation ($\% \pm \sigma$) and execution time (seconds) on MAG.

Methods	NMI	ARI	ACC	F1	Time
k-means	42.04	32.34	58.63	59.81	-
DGI	53.56 \pm 0.48	42.60 \pm 0.83	59.89 \pm 1.10	57.17 \pm 1.88	<u>36</u>
DMGI	49.71 \pm 1.37	38.91 \pm 1.35	53.57 \pm 0.54	49.59 \pm 1.39	118
HDMI	48.15 \pm 0.98	34.92 \pm 1.27	51.78 \pm 1.37	49.80 \pm 2.04	105
MGDCR	<u>54.43\pm1.17</u>	<u>43.98\pm1.16</u>	<u>61.37\pm2.46</u>	<u>60.53\pm3.19</u>	39
DMG	44.04 \pm 3.32	36.97 \pm 2.86	57.65 \pm 2.03	55.32 \pm 2.53	95
BMGC	57.01\pm0.19	47.84\pm0.27	65.31\pm1.25	63.68\pm1.84	25

dimension to 128 and the batch size to 5000. Table 3 presents the results with standard deviation and training time. Due to our scalable graph encoding that eliminates time-consuming neighbor sampling and graph convolution operations during training, BMGC achieves optimal results with the shortest training time.

In summary, across all datasets, BMGC consistently exhibits superior performance. The stable results obtained in these experiments demonstrate the effectiveness of our methods in addressing the view imbalance of multi-relational graphs.

6.6 Ablation Study

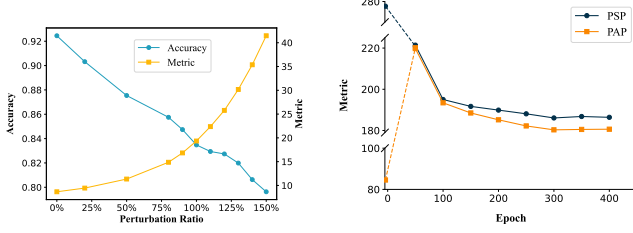
To validate the effectiveness of different components in our model, we compare the performance of BMGC with its three variants:

- Employing BMGC without \mathcal{L}_{ADV} to show the significance of alignment with the dominant view.
- Employing BMGC without \mathcal{L}_{ANF} to observe the impact of alignment with the node features.
- Employing BMGC without \mathcal{L}_{CLU} to reveal the influence of the dominant assignment on clustering performance.

Based on Table 4, we can draw the following conclusions. First, the results of BMGC are better than all variants, indicating that

Table 4: Performance of BMGC and its variants.

Variants	Amazon		ACM		DBLP		ACM2		Yelp	
	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC
BMGC	0.5768	0.7856	0.7841	0.9413	0.8013	0.9401	0.7285	0.9185	0.7173	0.9151
w/o \mathcal{L}_{ADV}	0.5303	0.7534	0.7366	0.9261	0.7773	0.9314	0.6054	0.8276	0.6763	0.8955
w/o \mathcal{L}_{ANF}	0.4234	0.6452	0.7667	0.9368	0.7923	0.9349	0.6762	0.8816	0.7041	0.9139
w/o \mathcal{L}_{CLU}	0.5629	0.7794	0.7726	0.9371	0.7857	0.9346	0.7263	0.7477	0.6864	0.9052



(a) The effectiveness of unsupervised dominant view mining. (b) The reliability of dynamic mining in the training process.

Figure 3: Case study on synthetic (left) and ACM (right) datasets. The specific meanings of the “Metric” in each figure can be found in Section 6.7.

all components are critical to our model. Second, the loss of alignment with the dominant view (\mathcal{L}_{ADV}) seems to make the most contribution to the results, while the loss for alignment with the node features (\mathcal{L}_{ANF}) contributes more to Amazon. This could be attributed to the universally subpar quality of all graph structures within the Amazon dataset, thereby amplifying the importance of node features. Additionally, the dominant assignment (\mathcal{L}_{CLU}) indeed improves clustering performance.

6.7 Case Study

Effectiveness of Unsupervised Mining. We delve deep into examining the impact of the perturbation ratio on both the accuracy and the unsupervised metric ($\|XX^T - X^v(X^v)^T\|_F^2 / N$) in View 3 of the synthetic dataset. As depicted in Fig. 3a, it is conspicuous that with the increase of the perturbation ratio from 0 to 150%, the accuracy consistently decreases, indicating a continuous decline in view quality. In parallel, the corresponding unsupervised metric indeed rises. This highlights the effectiveness of our unsupervised dominant view extraction method, aligning with the conclusions drawn in the theoretical analysis in Section 5.

Reliability of Dynamic Mining. To provide a detailed description of the process through which our model uncovers the dominant view, we demonstrate the evolution of the unsupervised metric ($\|XX^T - Z^v(Z^v)^T\|_F^2 / N$), used to mine the dominant view, for each view of the ACM dataset. As illustrated in Fig. 3b, the data points on the y-axis represent the aggregated node features used to initialize the dominant view. These points are connected by dashed lines to subsequent data points derived from node representations. Throughout the training, the metrics of both views decrease, and

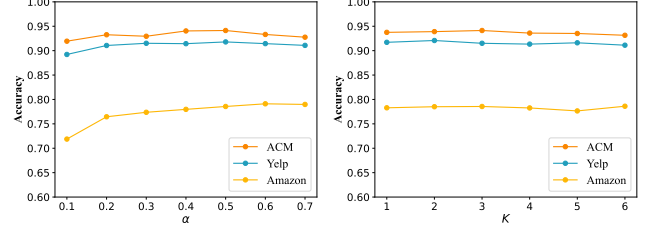


Figure 4: The influence of α (left) and K (right).

the PAP view consistently exhibits lower values compared to the PSP view. This consistent trend indicates that PAP emerges as the dominant view, aligning seamlessly with our empirical analysis. After 250 epochs, the metrics converge. This result emphasizes the reliability of dynamically excavating the dominant view.

6.8 Hyper-parameters Study

We conduct a hyper-parameter analysis on the teleport probability α and the radius of graph propagation K on three datasets ACM, Yelp, and Amazon. The result is given in Fig. 4. From the figure on the left, we can observe that our model shows low sensitivity to the change of α . However, when α is too low, the performance shows a noticeable decrease. This can be attributed to the fact that the lower value of α leads to a decreased influence of the features of the original nodes in the propagation process. In the figure on the right, we can see that the performance is stable to the change of K . Notable performance can be achieved when K is small, improving the efficiency of the model in practical applications.

7 Conclusion

In this study, we thoroughly investigate the prevalent challenge of view imbalance in real-world multi-relational graphs. We introduce a novel metric, the Aggregation Class Distance, to empirically quantify structural disparities among different graphs. To tackle view imbalance, we propose Balanced Multi-Relational Graph Clustering, which dynamically mines the dominant view throughout the training process, collaborating with representation learning to enhance clustering performance. Theoretical analysis validates the efficacy of unsupervised dominant view mining. Extensive experiments and in-depth analysis on both real-world and synthetic datasets consistently demonstrate the superiority of our model over existing state-of-the-art methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62276053).

References

- [1] Emmanuel Abbe. 2018. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research* 18, 177 (2018), 1–86.
- [2] Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. 2022. Effects of Graph Convolutions in Multi-layer Networks. In *The Eleventh International Conference on Learning Representations*.
- [3] Sudhanshu Chanpuriya and Cameron Musco. 2022. Simplified graph convolution with heterophily. *Advances in Neural Information Processing Systems* 35 (2022), 27184–27197.
- [4] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. 2020. Adaptive Universal Generalized PageRank Graph Neural Network. In *International Conference on Learning Representations*.
- [5] Shaohua Fan, Xiao Wang, Chuan Shi, Emiao Lu, Ken Lin, and Bai Wang. 2020. One2multi graph autoencoder for multi-view graph clustering. In *proceedings of the web conference 2020*. 3070–3076.
- [6] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. PMR: Prototypical Modal Rebalance for Multimodal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20029–20038.
- [7] Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. 2020. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of The Web Conference 2020*. 2331–2341.
- [8] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2022. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence* 45, 2 (2022), 2551–2566.
- [9] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.
- [10] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems* 34 (2021), 10944–10956.
- [11] Baoyu Jing, Chanyoung Park, and Hanghang Tong. 2021. Hdmi: High-order deep multiplex infomax. In *Proceedings of the Web Conference 2021*. 2414–2424.
- [12] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).
- [13] Zhiping Lin, Zhao Kang, Lizong Zhang, and Ling Tian. 2023. Multi-view attributed graph clustering. *IEEE Transactions on knowledge and data engineering* 35, 2 (2023), 1872–1880.
- [14] Yawen Ling, Jianpeng Chen, Yazhou Ren, Xiaorong Pu, Jie Xu, Xiaofeng Zhu, and Lifang He. 2023. Dual label-guided graph refinement for multi-view graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 8791–8798.
- [15] Liang Liu, Zhao Kang, Jijia Ruan, and Xixu He. 2022. Multilayer graph contrastive clustering network. *Information Sciences* 613 (2022), 256–267.
- [16] Xinwang Liu, Xinzong Zhu, Miaomiao Li, Lei Wang, Chang Tang, Jianping Yin, Dinggang Shen, Huaimin Wang, and Wen Gao. 2018. Late fusion incomplete multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence* 41, 10 (2018), 2410–2423.
- [17] Yue Liu, Wenxuan Tu, Sihang Zhou, Xinwang Liu, Linxuan Song, Xihong Yang, and En Zhu. 2022. Deep graph clustering via dual correlation reduction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 7603–7611.
- [18] Yuanfu Lu, Chuan Shi, Linmei Hu, and Zhiyuan Liu. 2019. Relation structure-aware heterogeneous information network embedding. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 4456–4463.
- [19] Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. 2022. Revisiting heterophily for graph neural networks. *Advances in neural information processing systems* 35 (2022), 1362–1375.
- [20] Yujie Mo, Yuhuan Chen, Yajie Lei, Liang Peng, Xiaoshuang Shi, Changan Yuan, and Xiaofeng Zhu. 2023. Multiplex Graph Representation Learning Via Dual Correlation Reduction. *IEEE Transactions on Knowledge and Data Engineering* (2023).
- [21] Yujie Mo, Yajie Lei, Jialie Shen, Xiaoshuang Shi, Heng Tao Shen, and Xiaofeng Zhu. 2023. Disentangled multiplex graph representation learning. In *International Conference on Machine Learning*. PMLR, 24983–25005.
- [22] Erlin Pan and Zhao Kang. 2021. Multi-view contrastive graph clustering. *Advances in neural information processing systems* 34 (2021), 2148–2159.
- [23] Erlin Pan and Zhao Kang. 2023. Beyond homophily: Reconstructing structure for graph-agnostic clustering. In *International Conference on Machine Learning*. PMLR, 26868–26877.
- [24] Erlin Pan and Zhao Kang. 2023. High-order multi-view clustering for generic data. *Information Fusion* 100 (2023), 101947.
- [25] Chanyoung Park, Donghyun Kim, Jiawei Han, and Hwanjo Yu. 2020. Unsupervised attributed multiplex network embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5371–5378.
- [26] Liang Peng, Xin Wang, and Xiaofeng Zhu. 2023. Unsupervised Multiplex Graph Learning with Complementary and Consistent Information. In *Proceedings of the 31st ACM International Conference on Multimedia*. 454–462.
- [27] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8238–8247.
- [28] Xiaowei Qian, Bingheng Li, and Zhao Kang. 2024. Upper Bounding Barlow Twins: A Novel Filter for Multi-Relational Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 14660–14668.
- [29] Meng Qu, Jian Tang, Jingbo Shang, Xiang Ren, Ming Zhang, and Jiawei Han. 2017. An attention-based collaboration framework for multi-view network representation learning. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*. 1767–1776.
- [30] Ylli Sadikaj, Justus Rass, Yllka Velaj, and Claudia Plant. 2023. Semi-Supervised Embedding of Attributed Multiplex Networks. In *Proceedings of the ACM Web Conference 2023*. 578–587.
- [31] Wenxuan Tu, Sihang Zhou, Xinwang Liu, Xifeng Guo, Zhiping Cai, En Zhu, and Jieren Cheng. 2021. Deep fusion clustering network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 9978–9987.
- [32] Petar Velicković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2018. Deep Graph Infomax. In *International Conference on Learning Representations*.
- [33] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413.
- [34] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12695–12705.
- [35] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The world wide web conference*. 2022–2032.
- [36] Zehong Wang, Qi Li, Donghua Yu, Xiaolong Han, Xiao-Zhi Gao, and Shigen Shen. 2023. Heterogeneous graph contrastive multi-view learning. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 136–144.
- [37] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. 2019. Simplifying graph convolutional networks. In *International conference on machine learning*. PMLR, 6861–6871.
- [38] Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, Bo Long, et al. 2023. Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning* 16, 2 (2023), 119–328.
- [39] Bin Zhang, Qianqiao Qiang, Fei Wang, and Feiping Nie. 2021. Flexible multi-view unsupervised graph embedding. *IEEE Transactions on Image Processing* 30 (2021), 4143–4156.
- [40] Changqing Zhang, Huazhu Fu, Qinghua Hu, Pengfei Zhu, and Xiaochun Cao. 2016. Flexible multi-view dimensionality co-reduction. *IEEE Transactions on Image Processing* 26, 2 (2016), 648–659.
- [41] Jianan Zhao, Xiao Wang, Chuan Shi, Zekuan Liu, and Yanfang Ye. 2020. Network schema preserving heterogeneous information network embedding. In *International joint conference on artificial intelligence (IJCAI)*.
- [42] Jiong Zhu, Ryan A Rossi, Anup Rao, Tung Mai, Nedin Lipka, Nesreen K Ahmed, and Danai Koutra. 2021. Graph neural networks with heterophily. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11168–11176.

A Algorithm

Algorithm 1 The pseudo-code of the proposed BMGC

Input: Node features X , adjacency matrices $\tilde{A}^1, \tilde{A}^2, \dots, \tilde{A}^V$ where V is the number of relations, the number of clusters C , initialized model parameters Θ .

- 1: Obtain view-specific aggregated features X^v by Eq. (3) before training;
- 2: Initialize the dominant view by Eq. (7);
- 3: **while** not reaching maximum epochs **do**
- 4: **if** every t epochs **then**
- 5: Recalculate the dominant view by Eq. (8);
- 6: **end if**
- 7: Obtain node representations Z^v with encoder f_Θ ;
- 8: Obtain the reconstruction of view-specific aggregated features \tilde{X}^v with decoder g_Θ ;
- 9: Calculate the reconstruction loss \mathcal{L}_{REC} by Eq. (6);
- 10: Calculate the co-aligned representation learning loss \mathcal{L}_{CAL} by Eq. (9), Eq. (10), Eq. (11) and Eq. (12);
- 11: Obtain the dominant assignment by applying k-means to the representations of the dominant view by Eq. (13);
- 12: Compute the soft alignment distribution Q^v and the target distribution P^v for each view by Eq. (14) and Eq. (15);
- 13: Obtain the final node representation $Z = [Z^1, \dots, Z^V]$ by concatenating representations from all views;
- 14: Apply k-means on the final representation Z and compute Q and P distribution of Z ;
- 15: Calculate the self-training clustering loss \mathcal{L}_{CLU} by Eq. (16);
- 16: Compute the overall objective \mathcal{L} by Eq. (17);
- 17: Back-propagate \mathcal{L} to update model weights;
- 18: **end while**
- 19: Apply k-means on the final node representation Z to obtain the clustering results;

Output: The final node representation Z and the clustering results.

B Detailed Proofs

Data Assumption. A multi-relational graph G has N nodes partitioned into 2 equally sized communities C_1 and C_2 . Let $c_1, c_2 \in \{0, 1\}^N$ be indicator vectors for membership in each community, that is, the j^{th} entry of c_i is 1 if the j^{th} node is in C_i and 0 otherwise. G has V views, each is generated by SBM [1], with intra- and inter-community edge probabilities p^v and q^v . G is such a graph model with a feature matrix $X = F + H \in \mathbb{R}^{N \times d_f}$, where each column of H follows a zero-centered, isotropic Gaussian noise distribution $\mathcal{N}(0, \sigma^2 I)$ and these columns are mutually independent. The matrix F is defined as $F = c_1 \mu_1^\top + c_2 \mu_2^\top$, where $\mu_1, \mu_2 \in \mathbb{R}^N$ has the same Euclidean norm $\|\mu\|$, representing the expected characteristic vector of each community. In addition, let $\bar{\mu} = \frac{1}{2}(\mu_1 + \mu_2)$ be the average of the feature vector means.

LEMMA 1. Let X^v be the aggregated feature matrix of the v -th view by applying SGC, with the number of hops K , to the expected adjacency matrix \tilde{A}^v and the feature matrix X . Then, $X^v = F^v +$

$c_1(\theta_1^v)^\top + c_2(\theta_2^v)^\top$, where $F^v = (\lambda_2^v)^K F + (1 - (\lambda_2^v)^K)(1\bar{\mu}^\top)$, θ_1^v and $\theta_2^v \in \mathbb{R}^{d_f}$ are both distributed according to $\mathcal{N}(0, \frac{1}{N}(1 + (\lambda_2^v)^{2K})\sigma^2 I)$, and $\lambda_2^v = \frac{p^v - q^v}{p^v + q^v} \in [-1, 1]$ is the second largest non-zero eigenvalue of the associated normalized adjacency matrix A^v .

THEOREM 1. Let \bar{X}_1^v and \bar{X}_2^v denote the centroid of aggregated features for each community in the v -th view. Then, $\mathbb{E}[\bar{X}_1^v - \bar{X}_2^v] = (\lambda_2^v)^K(\mu_1 - \mu_2)$ and $\mathbb{E}[XX^\top - X^v(X^v)^\top] = \frac{1 - (\lambda_2^v)^{2K}}{2}(\| \mu \|^2 - \mu_1^\top \mu_2) + (c_1 c_1^\top + c_2 c_2^\top - c_1 c_2^\top - c_2 c_1^\top) + \omega(\sigma^2)$, where $\omega(\sigma^2)$ represents the sum of terms containing σ^2 that are of negligible magnitude.

We first prove Lemma 1. Since the proof process is identical for each view, we omit the superscript v of some view-related symbols for simplicity in the proof.

PROOF OF LEMMA 1. In expectation, an entry \tilde{A}_{ij} of the adjacency matrix of the graph is p if both $i, j \in C_1$ or both $i, j \in C_2$, and it is q otherwise. The eigendecomposition $\mathbf{Q} \text{diag}(\lambda) \mathbf{Q}^\top$ of the associated normalized adjacency matrix A has two non-zero eigenvalues: $\lambda_1 = 1$, with eigenvector $\mathbf{q}_1 = \frac{1}{\sqrt{N}}\mathbf{1} = \frac{1}{\sqrt{N}}(c_1 + c_2)$, and $\lambda_2 = \frac{p-q}{p+q}$, with $\mathbf{q}_2 = \frac{1}{\sqrt{N}}(c_1 - c_2)$.

Zero-centered, isotropic Gaussian distributions are invariant to rotation, which means $\mathbf{Q}^\top H_{:,i} \sim \mathcal{N}(0, \sigma^2 I)$ for any orthonormal matrix \mathbf{Q} , so X^v can be expressed as follows:

$$\begin{aligned}
 X^v &= A^K X \\
 &= \mathbf{Q} \text{diag}(\lambda)^K \mathbf{Q}^\top (c_1 \mu_1^\top + c_2 \mu_2^\top + H) \\
 &= \mathbf{q}_1 \mathbf{q}_1^\top (c_1 \mu_1^\top + c_2 \mu_2^\top + H) + \lambda_2^K \mathbf{q}_2 \mathbf{q}_2^\top (c_1 \mu_1^\top + c_2 \mu_2^\top + H) \\
 &= \mathbf{q}_1 \left[\frac{\sqrt{N}}{2}(\mu_1^\top + \mu_2^\top) + h_1^\top \right] + \lambda_2^K \mathbf{q}_2 \left[\frac{\sqrt{N}}{2}(\mu_1^\top - \mu_2^\top) + h_2^\top \right] \\
 &= (c_1 + c_2) \left[\frac{1}{2}(\mu_1^\top + \mu_2^\top) + \frac{1}{\sqrt{N}} h_1^\top \right] \\
 &\quad + \lambda_2^K (c_1 - c_2) \left[\frac{1}{2}(\mu_1^\top - \mu_2^\top) + \frac{1}{\sqrt{N}} h_2^\top \right] \\
 &= \lambda_2^K (c_1 \mu_1^\top + c_2 \mu_2^\top) + \frac{(1 - \lambda_2^K)}{2} (c_1 + c_2)(\mu_1 + \mu_2)^\top + \\
 &\quad \frac{1}{\sqrt{N}} c_1 (h_1 + \lambda_2^K h_2)^\top + \frac{1}{\sqrt{N}} c_2 (h_1 - \lambda_2^K h_2)^\top \\
 &= \lambda_2^K F + (1 - \lambda_2^K) 1\bar{\mu}^\top + c_1 \theta_1^\top + c_2 \theta_2^\top
 \end{aligned} \tag{18}$$

PROOF OF THEOREM 1. Based on the Lemma 1, we can obtain the following.

$$\begin{aligned}
 \mathbb{E}[\bar{X}_1^v - \bar{X}_2^v] &= \mathbb{E} \left[\lambda_2^K \mu_1 + (1 - \lambda_2^K) \bar{\mu} + \theta_1 \right] \\
 &\quad - \mathbb{E} \left[\lambda_2^K \mu_2 + (1 - \lambda_2^K) \bar{\mu} + \theta_2 \right] \\
 &= \lambda_2^K (\mu_1 - \mu_2)
 \end{aligned} \tag{19}$$

Similarly, we can formulate the expression for $\mathbb{E}[XX^\top - X^v(X^v)^\top]$ as follows:

Table 5: Classification Accuracy (ACC), ACD, and Homophily Ratio (HR) across different ϕ values.

ϕ	-1	-0.5	-0.25	0	0.25	0.5	1
HR	0.0412	0.1748	0.3247	0.493	0.6832	0.8321	0.9599
ACD	0.3251	0.1606	0.0628	0.0274	0.0646	0.1645	0.3293
ACC	0.9829	0.8849	0.6637	0.5174	0.6711	0.8911	0.9863

$$\begin{aligned}
& \mathbb{E} [XX^\top - X^v(X^v)^\top] \\
&= \mathbb{E} [(F+H)(F+H)^\top] - \mathbb{E} \left[(\lambda_2^K F + (1-\lambda_2^K) \mathbf{1} \bar{\mu}^\top + c_1 \theta_1^\top \right. \\
&\quad \left. + c_2 \theta_2^\top) (\lambda_2^K F + (1-\lambda_2^K) \mathbf{1} \bar{\mu}^\top + c_1 \theta_1^\top + c_2 \theta_2^\top)^\top \right] \\
&= FF^\top + \mathbb{E}(HH^\top) - \mathbb{E} \left[\lambda_2^{2K} FF^\top + \lambda_2^K (1-\lambda_2^K) (F \bar{\mu} \mathbf{1}^\top + \mathbf{1} \bar{\mu}^\top F^\top) \right. \\
&\quad \left. + (1-\lambda_2^K)^2 \bar{\mu}^\top \bar{\mu} \mathbf{1} \mathbf{1}^\top + \theta_1^\top \theta_1 c_1 c_1^\top + \theta_2^\top \theta_2 c_2 c_2^\top \right. \\
&\quad \left. + \theta_1^\top \theta_2 (c_1 c_2^\top + c_2 c_1^\top) \right] \\
&= (1-\lambda_2^{2K}) FF^\top - \lambda_2^K (1-\lambda_2^K) \left[(c_1 \mu_1^\top + c_2 \mu_2^\top) \frac{1}{2} (\mu_1 + \mu_2) \mathbf{1}^\top \right. \\
&\quad \left. + \frac{1}{2} (\mu_1 + \mu_2)^\top (c_1 \mu_1^\top + c_2 \mu_2^\top) \right] - (1-\lambda_2^K)^2 \frac{1}{4} (\mu_1 + \mu_2)^\top \\
&\quad (\mu_1 + \mu_2) \mathbf{1} \mathbf{1}^\top + \mathbb{E}(HH^\top) - \mathbb{E} \left[\theta_1^\top \theta_1 c_1 c_1^\top + \theta_2^\top \theta_2 c_2 c_2^\top \right. \\
&\quad \left. + \theta_1^\top \theta_2 (c_1 c_2^\top + c_2 c_1^\top) \right] \\
&= (1-\lambda_2^{2K}) (c_1 \mu_1^\top + c_2 \mu_2^\top) (c_1 \mu_1^\top + c_2 \mu_2^\top)^\top - \frac{1}{2} (1-\lambda_2^{2K}) \\
&\quad (\|\mu\|^2 + \mu_1^\top \mu_2) \mathbf{1} \mathbf{1}^\top + d_f \sigma^2 I - \frac{d_f \sigma^2}{N} (1+\lambda_2^{2K}) (c_1 c_1^\top + c_2 c_2^\top) \\
&\quad - \mathbb{E} \left[\frac{1}{N} (h_1 + \lambda_2^K h_2)^\top (h_1 - \lambda_2^K h_2) \right] (c_1 c_2^\top + c_2 c_1^\top) \\
&= (1-\lambda_2^{2K}) (c_1 \mu_1^\top + c_2 \mu_2^\top) (c_1 \mu_1^\top + c_2 \mu_2^\top)^\top - \frac{1}{2} (1-\lambda_2^{2K}) \\
&\quad (\|\mu\|^2 + \mu_1^\top \mu_2) (c_1 + c_2) (c_1 + c_2)^\top + d_f \sigma^2 I - \frac{d_f \sigma^2}{N} (1+\lambda_2^{2K}) \\
&\quad (c_1 c_1^\top + c_2 c_2^\top) - \frac{d_f \sigma^2}{N} (1-\lambda_2^{2K}) (c_1 c_2^\top + c_2 c_1^\top) \\
&= \frac{1-\lambda_2^{2K}}{2} (\|\mu\|^2 - \mu_1^\top \mu_2) (c_1 c_1^\top + c_2 c_2^\top - c_1 c_2^\top - c_2 c_1^\top) \\
&\quad + d_f \sigma^2 I - \frac{d_f \sigma^2}{N} (1+\lambda_2^{2K}) (c_1 c_1^\top + c_2 c_2^\top) \\
&\quad - \frac{d_f \sigma^2}{N} (1-\lambda_2^{2K}) (c_1 c_2^\top + c_2 c_1^\top)
\end{aligned} \tag{20}$$

We posit that the magnitude of $\|\mu\|^2 - \mu_1^\top \mu_2$ far exceeds $d_f \sigma^2$. Otherwise, for example, when $d_f \sigma^2$ is notably large, the distinctions in the features of the nodes between different communities would be indiscernible. All node features would tend to converge toward random Gaussian noise, rendering the node features meaningless. Consequently, we use $\omega(\sigma^2)$ to denote the sum of terms containing σ^2 , highlighting that this term has a negligible magnitude in comparison. The overall expression is given by $\mathbb{E} [XX^\top - X^v(X^v)^\top] =$

$\frac{1-(\lambda_2^K)^{2K}}{2} (\|\mu\|^2 - \mu_1^\top \mu_2) (c_1 c_1^\top + c_2 c_2^\top - c_1 c_2^\top - c_2 c_1^\top) + \omega(\sigma^2)$. Therefore, we complete the proof.

C Synthetic Datasets

We propose the multi-relational cSBM to generate imbalanced multi-relational graphs. We first introduce the data generation process of cSBM. Here, N denotes the number of nodes, and all nodes are divided into two classes of equal size with node labels $y_i \in \{-1, +1\}$. Each node possesses a d_f -dimensional feature vector, obtained by random sampling from class-specific Gaussian distributions, as follows:

$$X_i = \sqrt{\frac{\mu}{N}} y_i u + \frac{H_i}{\sqrt{d_f}} \tag{21}$$

where $u \sim \mathcal{N}(0, I/d_f)$ and $H_i \in \mathbb{R}^{d_f}$ has independent standard normal entries. The adjacency matrix \tilde{A} of the generated cSBM graph is defined as:

$$\mathbb{P} [\tilde{A}_{ij} = 1] = \begin{cases} \frac{d + \lambda \sqrt{d}}{N} & \text{if } y_i y_j > 0 \\ \frac{d - \lambda \sqrt{d}}{N} & \text{otherwise.} \end{cases} \tag{22}$$

where d is the average degree of the generated graph. Note that λ and μ are hyper-parameters to control the proportion of contributions from the graph structure and node features, respectively.

In cSBM, the parameter $\phi = \frac{2}{\pi} \arctan \left(\frac{\lambda \sqrt{\xi}}{\mu} \right) \in [-1, 1]$ controls the degree of homophily, where $\xi = \frac{\mu}{f}$ is a control factor. A larger $|\phi|$ implies that the graph can provide stronger topological information, whereas when $\phi = 0$, only the node features are informative. The parameters of cSBM should satisfy the constraint $\lambda^2 + \frac{\mu^2}{\xi} = 1 + \epsilon$, $\epsilon > 0$ to generate informative graphs. We follow [4] for cSBM parameter settings, using $N = 5000$, $d_f = 2000$, $d = 5$, $\epsilon = 3.25$.

To generate multi-relational graphs, we first create a label indicator vector and feature matrix for nodes. We then use cSBM to generate V graphs with different structures but the same ϕ value ($\phi = 0.5$). Uniform values of ϕ ensure that the initial graph structures of each view have the same homophily ratio (0.825), further providing an equal level of topological information. We choose $V = 3$ to enhance realism. To simulate view imbalance among different views, we randomly add noisy edges to two of these graphs to induce perturbations, as mentioned in the main text.

D Additional Experiments

We add an experiment to demonstrate the effectiveness of ACD and its superior reliability compared to the existing supervised metric, the homophily ratio. We use the cSBM generative model to produce a series of graphs sharing the same node features and achieve various structures solely by adjusting the ϕ value. 30% of the nodes are randomly selected as the training set, with the rest forming the testing set. A linear layer is used as the classifier for view-specific features.

Table 5 clearly shows that ACD is positively correlated with classification accuracy, which empirically proves the effectiveness

Table 6: The statistics of the datasets.

Datasets	Nodes	Relation Types	Edges	Features	Classes
ACM	3,025	Paper-Subject-Paper (PSP)	2,210,761	1,870	3
		Paper-Author-Paper (PAP)	29,281		
DBLP	4,057	Author-Paper-Author (APA)	11,113	334	4
		Author-Paper-Conference-Paper-Author (APCPA)	5,000,495		
		Author-Paper-Term-Paper-Author (APTPA)	6,776,335		
ACM2	4,019	Paper-Subject-Paper (PSP)	4,338,213	1,902	3
		Paper-Author-Paper (PAP)	57,853		
Yelp	2,614	Business-User-Business(BUB)	528,332	82	3
		Business-Rating Level-Business (BLB)	1,487,306		
		Business-Service-Business (BSB)	2,477,722		
Amazon	7,621	Item-AlsoView-Item (IVI)	266,237	2,000	4
		Item-AlsoBought-Item (IBI)	1,104,257		
		Item-BoughtTogether-Item (ITI)	16,305		
MAG	113,919	Paper-Paper (PP)	1,806,596	128	4
		Paper-Author-Paper (PAP)	10,067,799		

of the ACD metric in measuring the relevance of graph structure to downstream tasks. Meanwhile, the existing supervised metric homophily ratio cannot adapt to various graph structures. For example, when $\phi = 1$ or -1 , both graph structures can achieve very high classification results, while the homophily ratio cannot maintain consistency. This demonstrates the superiority of ACD compared to the existing metric.

E Real-world Datasets

We employ five publicly available real-world benchmark datasets and a large-scale dataset. ACM [5], ACM2 [7], and DBLP [41] are citation networks. Yelp [18] and Amazon [9] are review networks. MAG [33] is a large-scale citation network. The statistics of these datasets are presented in Table 6.

- **ACM** contains 3,025 papers with graphs generated by two meta-paths (paper-subject-paper and paper-author-paper). The feature of each paper is a 1,870-dimensional bag-of-words representation of its abstract. Papers are categorized into three classes, *i.e.*, database, wireless communication, and data mining.
- **DBLP** contains 4,057 papers with graphs generated by three meta-paths (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.
- **ACM2** contains 4,019 papers with graphs generated by two meta-paths (paper-subject-paper and paper-author-paper). The feature of each paper is a 1,902-dimensional bag-of-words representation of its abstract. Papers are categorized into three classes, *i.e.*, database, wireless communication, and data mining.
- **Yelp** contains 2,614 businesses with graphs generated by three meta-paths (business-user-business, business-rating levels-business and business-service-business). The feature of each business is an 82-dimensional bag-of-words representation of its rating information. Businesses are categorized into three classes, *i.e.*, Mexican flavor, hamburger type, and food bar.

- **Amazon** contains 7,621 items with graphs generated by three meta-paths (item-alsoView-item, item-alsoBought-item and item-boughtTogether-item). The feature of each item is a 2,000-dimensional bag-of-words representation of its description. Items are categorized into four classes, *i.e.*, beauty, automotive, patio lawn and garden, and baby.
- **MAG** is a subset of OGBN-MAG [33], consisting of the four largest classes. MAG contains 113,919 papers with graphs generated by two meta-paths (paper-author-paper and paper-paper). Each paper is associated with a 128-dimensional word2vec feature vector.

F Baselines

In this section, we give brief introductions of the baselines which are not described in the main paper due to the space constraint.

- **HAN** [35]: HAN is a supervised multiview graph learning method that performs multiview fusion through node-level attention and semantic-level attention, and utilizes node labels for training.
- **VGA**E [12]: VGA is a graph autoencoder that learns node embeddings via variational autoencoders. Both the encoder and the decoder are implemented with the graph convolutional network.
- **DGI** [32]: DGI maximizes the mutual information between patch representations and corresponding high-level summaries of graphs which are derived using the graph convolutional network.
- **O2MAC** [5]: O2MAC initially selects the most informative view as input and optimizes by reconstructing the graph structures of all views, coupled with self-supervised clustering loss.
- **DMGI** [25]: DMGI is an unsupervised multiplex network embedding method integrating the node embeddings from multiple graphs by minimizing the disagreements among the view-specific node embeddings and utilizing a universal discriminator.
- **MvAGC** [13]: MvAGC is a multiview attributed graph clustering method that utilizes a graph filter, the selection of anchor points, and a novel regularizer for high-order neighborhood information exploration.
- **HDMI** [11]: HDMI is a self-supervised model for learning node embedding on multiplex networks. It designs a joint supervision signal and combines node embedding from different layers by an attention-based fusion module.
- **MCGC** [22]: MCGC is a multiview contrastive graph clustering method, which contains graph filtering, graph learning, and graph contrastive components. And it learns a new consensus graph rather than the initial graph.
- **MGDCR** [20]: MGDCR is a self-supervised multiplex graph representation learning method, which jointly mines the common and private information in the multiplex graph while minimizing the redundant information within node representations using Barlow Twins loss.
- **DualGR** [14]: DualGR is a multiview graph clustering approach aimed at low homophilous graphs, which contains

a dual label-guided graph refinement module and a graph encoder module.

- **DMG** [21]: DMG is an unsupervised multiplex graph representation learning method that disentangles comprehensive and clear common information while capturing more complementary and less noisy private information.
- **BTGF** [28]: BTGF is a multi-relational clustering method with a novel graph filter motivated by the theoretical analysis of Barlow Twins, which makes the inner product positive semi-definite to upper bound Barlow Twins.