DEEP FAIR GRAPH CLUSTERING VIA ADVERSARIAL CLUSTERING STRUCTURE DEBIASING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep fair graph clustering (DFGC) aims to learn unbiased graph representations that prevent clustering results from being dominated by specific sensitive groups. Existing methods typically mitigate sensitive-attribute bias by training an adversarial discriminator to predict sensitive attribute labels from node representations. However, this node-wise prediction objective overlooks the sensitive-group clustering structure encoded by graph topology. Consequently, such structural bias may persist in the learned node representations, yielding clusters that are overly aligned with sensitive-group boundaries. To address this issue, we propose a novel DFGC framework termed **A**dversarial **C**lustering **S**tructure **D**ebiasing **(ACSD)**, which learns fair graph representations by adversarially debiasing sensitive clustering structure. Specifically, a sensitive adversarial encoder is designed to extract sensitive clustering structure through modularity maximization and sensitive cluster-membership reconstruction. The clustering backbone is then optimized to hinder the extraction of sensitive clustering structures while preserving clustering quality, promoting graph clustering embeddings that are less dependent on sensitive-group structures. Extensive experiments on benchmarks demonstrate the effectiveness and superiority of ACSD over existing state-of-the-art methods.

1 INTRODUCTION

Deep graph clustering (DGC) integrates node attributes and graph topology to learn representations that reveal latent clustering structures, and has achieved remarkable progress in recent years (Liang et al., 2024; Tu et al., 2024; 2025). Alongside advances in clustering performance, growing concerns over algorithmic bias have motivated increasing interest in the fairness of DGC (Ghodsi et al., 2024; Zhou et al., 2024; Parraga et al., 2025). Beyond producing accurate clusters, deep fair graph clustering (DFGC) seeks to prevent sensitive attributes, such as gender or race, from adversely influencing cluster assignments (Zhang et al., 2026). This objective is particularly challenging for graph data. Unlike images and text (Li et al., 2020; Zeng et al., 2023), whose samples are often modeled independently, graphs explicitly encode dependencies among nodes through their connections. Sensitive information may therefore be embedded not only in node attributes but also in the relational structure of the graph. Effective DFGC consequently requires learned representations that preserve task-relevant clustering structure while suppressing sensitive information carried by both node attributes and graph topology.

Most existing DFGC methods usually learn fair representations through adversarial learning (Zhang et al., 2026). A typical approach employs a discriminator to predict the sensitive attributes of each node from its representations, while training the graph encoder to prevent such predictions (Wang et al., 2026). Although this strategy promotes node-level sensitive invariance, it does not explicitly account for the relational geometry of the representation space. In particular, the discriminator evaluates whether sensitive labels can be inferred from individual node representations, whereas graph clustering groups nodes according to their relative affinity. These affinities are further shaped by graph message passing, which couples the representations of connected nodes (Hu et al., 2025). When graph connectivity is associated with sensitive attributes (Newman, 2003), nodes from the same sensitive group may remain close in the representation space even if their sensitive labels are difficult to predict individually. Consequently, sensitive-group structure can persist in pairwise relationships and induce clusters that are disproportionately aligned with sensitive-group boundaries. Node-wise sensitive invariance alone is therefore insufficient to guarantee fair graph clustering.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This objective mismatch raises a fundamental question: *how can we explicitly mitigate sensitive-group clustering structure in the learned graph representations while promoting a fair distribution of sensitive groups across the resulting clusters?* Answering this question requires considering how such structure emerges. DGC methods commonly employ graph convolutional networks (GCNs) to jointly encode node attributes and graph structure, increasing the similarity between representations of connected nodes (Liu et al., 2022b). Since nodes within the same sensitive group are more likely to be connected in graphs, message passing can amplify the similarity among nodes sharing the same sensitive attribute and thereby preserve sensitive communities in the embedding space (Tu et al., 2023). Moreover, clustering depends on the global geometry of the representation space, where cluster assignments are determined collectively by inter-node affinities rather than by individual representations in isolation. Sensitive clustering structure is therefore reflected in two complementary forms: topology-consistent community organization and pairwise sensitive-group co-membership. Effectively mitigating such structural bias requires an adversary that jointly captures both forms, beyond conventional discriminators limited to node-wise sensitive-attribute prediction.

To this end, we propose a novel deep fair graph clustering method termed **A**dversarial **C**lustering **S**tructure **D**ebiasing **(ACSD)**, a simple yet effective framework learning fair graph representations by adversarially debiasing sensitive clustering structure. Specifically, an adversarial sensitive-structure encoder maps node representations into a space for extracting sensitive clustering structure, where multiple prototypes for each sensitive group are used to derive soft group assignments. Subsequently, the adversary maximizes modularity to capture community structure associated with sensitive groups based on the soft assignments. Additionally, it reconstructs pairwise sensitive-group co-membership from these assignments, encouraging the extracted clustering structure to reflect which nodes belong to the same sensitive group. These objectives jointly encourage the sensitive-structure encoder to encode graph communities and sensitive grouping relations. The clustering backbone is trained to oppose these joint objectives, thereby reducing sensitive bias in the learned node representations. Our main contributions are summarized as follows:

- **New perspective.** We examine fairness in graph clustering from the perspective of sensitive clustering structure, highlighting that node-wise sensitive invariance alone does not guarantee the elimination of sensitive grouping information from representation affinities.
- **Novel framework.** We propose ACSD, which introduces a structural adversarial mechanism between a sensitive-structure encoder and a clustering backbone. The adversary identifies sensitive clustering structure, while the clustering backbone hinders this extraction to reduce sensitive bias in node representations and promote fair cluster assignments.
- **Superior performance.** Extensive experiments on benchmark datasets demonstrate the superiority of ACSD in improving fairness compared with state-of-the-art methods.

2 Problem Setup and Motivation

Problem Setup Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ denote an undirected attributed graph with n nodes, where $\mathcal{V}$ and $\mathcal{E}$ are the node and edge sets, respectively, and $\mathbf{X} \in \mathbb{R}^{n \times d_x}$ is the node attribute matrix. The graph topology is represented by the adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$, where $\mathbf{A}_{ij} = 1$ if and only if $(i,j) \in \mathcal{E}$. In fair graph clustering, each node i has a sensitive label $s_i \in \{1, \ldots, G\}$. Let $\mathcal{V}_g = \{i \in \mathcal{V} : s_i = g\}$ and $n_g = |\mathcal{V}_g|$ denote the node set and size of sensitive group g, respectively. Deep graph clustering learns a unified representation from node attributes and graph topology:

$$\mathbf{H} = f_g(\mathbf{X}, \mathbf{A}; \theta_g) \in \mathbb{R}^{n \times d_h}, \tag{1}$$

where f_g denotes the graph encoder parameterized by θ_g, which maps node attributes and graph topology to node representations, and the i-th row of $\mathbf{H}$ is $\mathbf{h}_i^\top$. Clustering is then performed on the learned node representations to partition the nodes into K clusters. Fair graph clustering aims to preserve clustering quality while preventing the resulting clusters from being disproportionately dominated by particular sensitive groups, encouraging the sensitive-group proportions within each cluster to match those of the overall population (Chierichetti et al., 2017).

Motivation for Clustering Structure Adversary Graph topology plays a critical role in shaping the representation space used for clustering. Through neighborhood aggregation, graph encoders can

increase the similarity of representations within graph communities, allowing the learned embeddings to reflect the community structure of the original graph (Yang et al., 2024). When connections are more prevalent within sensitive groups, this process can also reinforce affinities among nodes sharing sensitive attributes (Newman, 2003). Consequently, the learned representations may form clusters aligned with sensitive groups, leading to imbalanced sensitive-group proportions within the resulting clusters.

Clustering determines clusters by considering the overall organization of node representations and the relationships among them, whereas sensitive-label adversarial learning evaluates the predictability of individual sensitive labels. To illustrate this distinction, consider two distinct nodes i, j:

$$r_{ij}^s = \mathbb{I}(s_i = s_j), \qquad w_{ij} = \sigma(\mathbf{h}_i, \mathbf{h}_j), \tag{2}$$

where r_{ij}^s indicates whether nodes i and j belong to the same sensitive group, and $\sigma(\cdot, \cdot)$ is a pairwise affinity function measuring similarity between node representations for clustering. Sensitive-label prediction assesses whether $\mathbf{h}_i$ reveals s_i, while a relational assessment examines whether w_{ij} provides information about r_{ij}^s. Since cluster structure can be characterized by which pairs of nodes belong together, identifying these group relations provides a way to assess sensitive cluster structure beyond individual label predictability. The theoretical discussion in Section 3.5 shows that node-wise independence does not, in general, imply independence between representation affinities and sensitive same-group relations, that is:

$$I(\mathbf{h}_i; s_i) = 0,\ \forall i \quad \not\Rightarrow \quad I(w_{ij}; r_{ij}^s) = 0, \qquad i \neq j. \tag{3}$$

This limitation motivates using sensitive clustering structure as an explicit adversarial target, aligning the debiasing objective with graph clustering to reduce sensitive grouping information in the learned representations. Beyond sensitive same-group relations, we also consider how sensitive groupings align with the global community structure of the graph. Pairwise relations characterize which nodes belong to the same sensitive group, but do not explicitly assess whether these groupings form communities supported by graph connectivity. We therefore introduce a modularity-based adversarial objective to complement the pairwise criterion, jointly mitigating sensitive-group co-membership and sensitive community structure in the learned representations.

3 METHOD

Based on above analysis, we introduce an adversarial sensitive-structure encoder that aims to extract sensitive clustering structure from the node representations learned by the clustering backbone. The encoder jointly maximizes modularity and reconstructs sensitive-group co-membership, encouraging the identified structure to capture graph community and align with sensitive-group membership. The clustering backbone is trained to oppose this joint objective while optimizing its original clustering objective. The overall framework is illustrated in Figure 1.

3.1 SENSITIVE CLUSTER STRUCTURE INDUCTION

We instantiate the graph encoder in Equation 1 with a single-layer GCN to obtain node representations $\mathbf{H}$ (Devvrit et al., 2022). We then introduce an MLP-based sensitive-structure encoder parameterized by θ_s to extract sensitive cluster structure from the learned node representations:

$$\mathbf{H}^s = f_s(\mathbf{H}; \theta_s) \in \mathbb{R}^{n \times d_s}. \tag{4}$$

To enable differentiable learning of sensitive cluster structure, we use group-specific prototypes to compute soft sensitive-group assignments for each node. For each sensitive group g, we apply k-means to its sensitive embeddings $\{\mathbf{h}_i^s : i \in \mathcal{V}_g\}$ and use the resulting m cluster centers as prototypes, denoted by $\{\mathbf{p}_{g\ell}\}_{\ell=1}^m$. We then compute the soft assignment of each node to each sensitive group by aggregating its affinities to the prototypes of that group:

$$q_{ig} = \frac{\sum_{\ell=1}^m \exp\big(-\|\mathbf{h}_i^s - \mathbf{p}_{g\ell}\|_2^2/\tau\big)}{\sum_{g'=1}^G \sum_{\ell=1}^m \exp(-\|\mathbf{h}_i^s - \mathbf{p}_{g'\ell}\|_2^2/\tau)}, \tag{5}$$

where τ is a temperature parameter controlling assignment sharpness. The resulting assignment matrix is $\mathbf{Q} = [q_{ig}] \in [0, 1]^{n \times G}$, with $\sum_{g=1}^G q_{ig} = 1$ for every node i. Thus, each column corresponds

to a sensitive group, aggregating the contributions of its multiple prototypes. The soft assignment matrix $\mathbf{Q}$ serves as the basis for the subsequent modularity and sensitive-group co-membership objectives, enabling differentiable learning of sensitive clustering structure.

3.2 Sensitive Structure Adversary

We train the sensitive adversary to extract sensitive cluster structure through two complementary objectives. Modularity maximization encourages the extracted structure to reflect graph communities at the global level, while pairwise sensitive relation reconstruction captures sensitive grouping patterns by predicting whether pairs of nodes belong to the same sensitive group. These objectives guide the adversary to extract sensitive cluster structure through both global community organization and pairwise grouping relations.

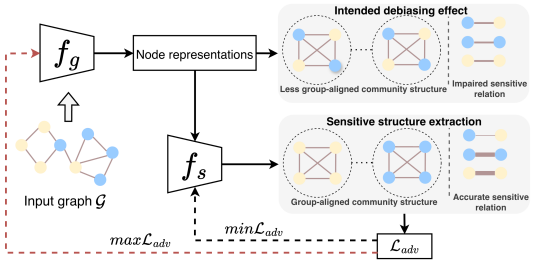

Figure 1: Illustration of the structural adversarial mechanism of ACSD. The sensitive encoder f_s jointly captures sensitive clustering structure through modularity maximization and pairwise sensitive relations reconstruction. With sensitive clustering structure as the adversarial objective, the graph encoder f_g is trained to hinder its extraction in order to reduce bias in node representations. Node colors indicate sensitive groups, and line thickness in the node-pair diagrams indicates predicted same-group affinity.

Global Community Structure via Modularity Graph encoders can incorporate community structure into the learned representations. We therefore encourage the sensitive adversary to extract group assignments that exhibit community structure in the original graph. To evaluate this global organization, we adopt modularity to assess the community structure associated with sensitive groups based on the soft assignment matrix $\mathbf{Q}$.

Let $\mathbf{A}$ denote the original undirected adjacency matrix without self-loops. Let:

$$\mathbf{B}_{\text{mod}} = \mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\top}{\nu}, \qquad \mathbf{d} = \mathbf{A}\mathbf{1}_n, \qquad \nu = \mathbf{1}_n^\top, \tag{6}$$

where $\mathbf{B}_{\text{mod}}$ denotes the modularity matrix (Tsitsulin et al., 2023), $\mathbf{d}$ is the node degree vector, and ν is twice the number of edges for an unweighted graph. Given the sensitive-group assignment matrix $\mathbf{Q}$, we compute the modularity score of the soft sensitive-group assignments:

$$\mathcal{M}_s(\mathbf{Q}; \mathbf{A}) = \frac{1}{\nu} \operatorname{Tr}\left(\mathbf{Q}^\top \mathbf{B}_{\text{mod}} \mathbf{Q}\right). \tag{7}$$

Maximizing this score encourages the groups inferred from the sensitive embeddings to exhibit stronger community structure in the original graph, thereby assessing the extracted structure at the community level. The sensitive adversary then maximizes modularity by minimizing:

$$\mathcal{L}_{\text{mod}} = -\mathcal{M}_s(\mathbf{Q}; \mathbf{A}). \tag{8}$$

Pairwise Sensitive Relation Reconstruction We then characterize sensitive cluster structure through sensitive-group co-membership to indicate which nodes belong to the same sensitive group. Unlike individual sensitive-label prediction, this objective evaluates grouping relations using pairs of node representations. It also ensures that the extracted grouping is supervised by sensitive attributes, since community structure alone does not identify sensitive groups.

For distinct nodes i, j, the ground-truth relation is $r^s_{ij} = \mathbb{I}(s_i = s_j)$. We predict this relation from the inner product of their sensitive-group assignments:

$$\widehat{r}^s_{ij} = \mathbf{q}_i^\top \mathbf{q}_j = \sum_{g=1}^{G} q_{ig} q_{jg}. \tag{9}$$

Since $\mathbf{q}_i$ and $\mathbf{q}_j$ are probability vectors, $\widehat{r}^s_{ij} \in [0, 1]$ measures their soft same-group agreement and is used as the predicted probability in the binary cross-entropy loss. Same-group pairs are

encouraged to have consistent assignments, whereas different-group pairs are encouraged to have distinct assignments.

We sample node pairs to avoid the quadratic computational cost of evaluating all pairwise relations. Let $\mathcal{B}$ denote a sampled batch containing equal numbers of same-group and different-group node pairs, excluding self-pairs. We then minimize the binary cross-entropy:

$$\mathcal{L}_{\mathrm{rel}} = -\frac{1}{|\mathcal{B}|} \sum_{(i,j)\in\mathcal{B}} \left[r_{ij}^{s} \log \widehat{r}_{ij}^{s} + (1 - r_{ij}^{s}) \log(1 - \widehat{r}_{ij}^{s}) \right]. \tag{10}$$

Here, the ground-truth relation are determined by sensitive-group membership, irrespective of whether the two nodes share an edge. The objective therefore assesses sensitive grouping relations rather than reconstructing graph adjacency.

Sensitive Structure Adversary Optimization With the clustering backbone fixed, we train the sensitive-structure encoder f_s by minimizing:

$$\min_{\theta_s} \ \mathcal{L}_{\mathrm{adv}} = \mathcal{L}_{\mathrm{rel}} + \alpha \mathcal{L}_{\mathrm{mod}}, \tag{11}$$

where α denote the hyperparameter to balance the two objectives. Relation reconstruction captures sensitive grouping patterns through pairwise membership supervision, while modularity evaluates their global community organization in the graph. They train the adversary to extract sensitive cluster structure that is consistent with sensitive attributes and supported by graph connectivity. The clustering backbone is subsequently trained to oppose this joint extraction objective.

3.3 Cluster-level Fairness Regularization

To explicitly encourage proportional fairness, we align the sensitive-group distribution within each task cluster with the population distribution. Let c_{ik} denote the soft assignment of node i to task cluster k, computed by applying softmax to the negative squared distances between backbone embeddings and k-means centers. The proportion of sensitive group g in cluster k is

$$b_{kg} = \frac{\sum_{i\in\mathcal{V}_g} c_{ik}}{\sum_{i=1}^{n} c_{ik}}. \tag{12}$$

With population proportions $\pi_g = n_g/n$, we encourage the sensitive-group distribution within each cluster to match the population distribution by minimizing their Kullback–Leibler divergence:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{g=1}^{G} \pi_g \log \frac{\pi_g}{b_{kg}}, \tag{13}$$

where K is the number of task clusters. Each cluster receives equal weight, and gradients propagate through the soft assignments to the backbone embeddings.

3.4 Adversarial Debiasing and Optimization

The overall framework is optimized by the following min–max adversarial objective:

$$\min_{\theta_g} \max_{\theta_s} \ \mathcal{L}_{\mathrm{clu}} - \beta \mathcal{L}_{adv} + \gamma \mathcal{L}_{\mathrm{KL}}, \tag{14}$$

where θ_g and θ_s denote the parameters of the clustering backbone and sensitive encoder, respectively. The coefficients β and γ denote the hyperparameters to balance the weight of each component. Following S3GC (Devvrit et al., 2022), we adopt a contrastive loss as the clustering objective $\mathcal{L}_{\mathrm{clu}}$ for the backbone network. Maximizing this adversarial objective with respect to θ_s minimizes the sensitive relation and modularity losses, training the adversary to extract sensitive clustering structure. Minimizing it with respect to θ_g opposes this extraction while optimizing clustering quality and proportional fairness.

We alternately update the two encoders. Backbone embeddings are detached during adversary updates. During backbone updates, the adversary parameters are frozen, but gradients propagate through the adversary to the backbone. Finally, we discard the sensitive adversary and apply k-means to the learned node representations to obtain the final clustering results. The detailed procedure of ACSD is provided in Algorithm 1.

Algorithm 1 The training procedure of ACSD

Input: Graph $\mathcal{G}$; number of clusters K; epochs T.

1: Initialize the graph encoder f_g and sensitive encoder f_s.
2: **for** $e = 1, \ldots, T$ **do**
3: Obtain node representations $\mathbf{H}$ by Eq. 1.
4: Obtain sensitive representations $\mathbf{H}^s$ by Eq. 4.
5: Obtain node soft assignments $\mathbf{Q}$ by Eq. 5.
6: Update θ_s by minimizing $\mathcal{L}_{\text{adv}}$ with $\mathbf{H}$ detached.
7: Recompute the losses with θ_s fixed and gradients retained to the backbone.
8: Update θ_g by minimizing $\mathcal{L}_{\text{clu}} - \beta\mathcal{L}_{\text{adv}} + \gamma\mathcal{L}_{\text{KL}}$.
9: **end for**
10: Obtain the final representations $\mathbf{H}$ and apply k-means with K clusters.

Output: Cluster assignments.

3.5 Theoretical Discussion

We first examine whether independence between individual node representations and their sensitive labels is sufficient to eliminate sensitive cluster structure from representation affinities. We treat the labels and representations as random variables under a joint distribution. For this analysis, we specialize the affinity function to the inner product, $w_{ij} = \mathbf{h}_i^\top \mathbf{h}_j$, and retain $r_{ij}^s = \mathbb{I}(s_i = s_j)$.

Theorem 1 (Node-wise invariance does not imply affinity invariance) *For any $G \geq 2$ and group sizes $n_1, \ldots, n_G \geq 2$ with $n = \sum_{g=1}^{G} n_g$, there exists a joint distribution of sensitive labels and representations $\mathbf{h}_i \in \mathbb{R}^G$, containing exactly n_g nodes in group g, such that:*

$$I(\mathbf{h}_i; s_i) = 0, \qquad \forall i, \tag{15}$$

while, for every distinct pair i, j,

$$I(w_{ij}; r_{ij}^s) = \mathsf{H}(r_{ij}^s) > 0, \tag{16}$$

where $\mathsf{H}(\cdot)$ denotes entropy.

Theorem 1 shows that complete sensitive same-group information can remain in pairwise affinities even when individual representations are independent of their sensitive labels. In the construction, $w_{ij} = r_{ij}^s$, so the affinity matrix directly identifies which nodes belong to the same sensitive group. It establishes a limitation of node-wise marginal independence, rather than a claim that every sensitive-label adversary fails.

Corollary 1 (Extension to group-dependent graph connectivity) *For any $0 \leq p_{\text{out}} < p_{\text{in}} \leq 1$, the joint distribution in Theorem 1 can be extended to an undirected graph satisfying:*

$$\Pr(A_{ij} = 1 \mid r_{ij}^s = 1) = p_{\text{in}}, \quad \Pr(A_{ij} = 1 \mid r_{ij}^s = 0) = p_{\text{out}}, \tag{17}$$

while preserving both conclusions of the theorem. Moreover, $I(A_{ij}; r_{ij}^s) > 0$ for every $i \neq j$.

This indicates that the same limitation can also occur in graphs where connections are more prevalent within sensitive groups. Since real-world graphs commonly exhibit pronounced assortativity (Newman, 2003), where nodes sharing similar sensitive attributes are more likely to be connected and form densely associated communities. Proofs are provided in Appendix. These results show that node-wise independence alone does not guarantee the elimination of sensitive grouping information from representation affinities, and that this limitation can also arise when graph connectivity is associated with sensitive groups. This provides theoretical motivation for using sensitive clustering structure as an explicit adversarial objective in fair graph clustering.

4 Experiment

Table 2: Clustering performance comparison between classical deep graph clustering methods and fair graph clustering methods. For each dataset, we report ACC (%), NMI (%), and Balance (BAL). ACSD achieves strong fairness performance across all six datasets while maintaining competitive clustering quality overall. OOM indicates that the method runs out of memory.

Dataset	Metric	Classical Deep Graph Clustering					Fair Graph Clustering			
		SDCN	MVGRL	DFCN	MAGI	MHGC	FNM	iFairNMTF	FairGC	**Ours**
New Bail	ACC	62.36	OOM	64.27	68.95	58.84	62.35	50.86	68.32	**69.34**
	NMI	0.01		2.29	10.39	8.05	0.01	0.04	**10.47**	8.41
	BAL	0.000		0.716	0.828	0.839	0.000	0.968	0.783	**0.974**
New German	ACC	**69.90**	66.20	64.30	53.20	53.00	53.40	55.27	52.93	68.30
	NMI	0.12	1.93	0.63	**3.73**	0.00	0.00	1.88	0.17	1.89
	BAL	0.000	0.315	0.285	0.405	0.193	0.286	0.422	0.346	**0.435**
New Credit	ACC	63.66	OOM	OOM	53.42	OOM	**77.90**	50.38	53.06	66.51
	NMI	1.02			1.35		0.09	0.01	1.30	**1.69**
	BAL	0.030			0.058		0.000	**0.098**	0.061	0.093
Occupation	ACC	66.32	54.91	52.65	54.00	60.30	63.46	53.65	61.63	**67.61**
	NMI	0.14	0.30	0.42	1.14	0.01	0.97	0.03	4.77	**9.56**
	BAL	0.791	0.764	0.736	0.682	0.824	0.966	0.970	0.640	**0.985**
Sport	ACC	51.77	50.11	51.88	97.23	**99.48**	50.40	50.40	53.48	59.66
	NMI	0.73	0.00	0.30	83.64	**95.61**	0.07	0.01	0.64	4.82
	BAL	0.484	0.897	0.727	0.073	0.073	0.600	0.901	0.839	**0.981**
Syn-1	ACC	58.08	76.04	62.16	74.88	75.14	81.40	50.82	60.71	**82.36**
	NMI	1.89	20.55	4.29	18.68	20.93	31.06	0.03	5.48	**32.78**
	BAL	0.773	0.390	0.628	0.378	0.231	0.484	0.957	0.595	**0.969**

Experimental Setup To evaluate the proposed ACSD on fair graph clustering, we use six graph datasets (Qian et al., 2024), namely New Bail, New German, New Credit, Occupation, Syn-1, and Sport. These datasets are extensively adopted in fair graph learning. The dataset statistics are summarized in Table 1, more details about the datasets are provided in the Appendix.

Table 1: Dataset statistics.

Dataset	# of nodes	Sensitive attribute	Clusters	Balance
New Bail	18876	Race(2)	2	0.974
New German	1000	Gender(2)	2	0.449
New Credit	30000	Age(2)	2	0.098
Occupation	6951	Gender(2)	2	0.985
Sport	3508	Race(2)	2	0.988
Syn-1	5000	0/1	2	0.970

We compare ACSD with three categories of baselines: (1) conventional deep graph clustering baselines, including SDCN (Bo et al., 2020), MVGRL (Zhu et al., 2021), DFCN (Tu et al., 2021), MAGI (Liu et al., 2024), and MHGC (Ren et al., 2025), which do not explicitly incorporate fairness objectives; (2) non-deep fair graph clustering methods, including FNM (Li et al., 2023) and iFairNMTF (Ghodsi et al., 2024), which incorporate fairness into spectral partitioning and matrix factorization, respectively; and (3) deep fair graph clustering methods, represented by FairGC (Zhang et al., 2026), which integrates fairness objectives into neural representation learning.

Implementation Details For simplicity, we use a single-layer GCN encoder and a contrastive clustering objective. The sensitive-structure encoder is implemented as a two-layer MLP. For each sensitive group, we construct 10 prototypes by applying k-means to the corresponding sensitive representations. We adopt Accuracy (ACC) and Normalized Mutual Information (NMI) to evaluate clustering utility. To assess fairness, we report Balance (Chierichetti et al., 2017), defined as the worst-case (minimum) group balance across clusters. Let $s \in \{0, 1\}$ denote the sensitive attribute, $c \in \{1, \ldots, K\}$ the cluster index, and $n_{c,s}$ the number of samples in cluster c with sensitive value s. The Balance is then defined as:

$$\text{Balance} = \min_{c \in \{1,\ldots,K\}} \frac{\min(n_{c,0},\, n_{c,1})}{\max(n_{c,0},\, n_{c,1})} \in [0, 1]. \tag{18}$$

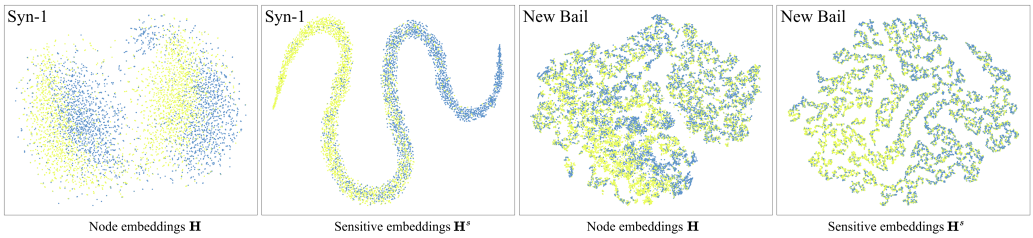

Figure 2: t-SNE visualizations of the node representations $\mathbf{H}$ and sensitive representations $\mathbf{H}^s$ on the New Bail and Syn-1 datasets. Colors indicate sensitive groups. Nodes from the same sensitive group exhibit clearer grouping in $\mathbf{H}^s$, whereas different sensitive groups are more mixed in $\mathbf{H}$.

A larger Balance metric indicates more balanced group representation within clusters, thus representing lower bias.

4.1 Performance Comparison Results

As shown in Table 2, we report the node clustering performance of eight methods across six datasets. From these results, three major observations can be drawn: *(1) Overall fairness.* Overall, ACSD achieves the highest BAL scores on five of the six datasets and the second-highest score on New Credit. For example, on Occupation and Syn-1, ACSD achieves Balance scores of 0.985 and 0.969, respectively, outperforming the second-best method, iFairNMTF, by 0.015 and 0.012. In particular, it achieves near-perfect group balance on New Bail, Occupation, and Syn-1, with BAL scores of 0.974, 0.985, and 0.969, respectively. These consistent results demonstrate the effectiveness of ACSD in improving clustering fairness. *(2) Comparison with fair graph clustering methods.* ACSD outperforms both non-deep fair graph clustering methods and the deep fair graph clustering baseline in terms of Balance. While these methods already incorporate fairness considerations, ACSD further improves clustering fairness by introducing an explicit adversarial objective for sensitive cluster structure. *(3) Comparison with conventional graph clustering methods.* Although some conventional methods achieve higher ACC, their Balance scores are substantially lower. For example, on Sport, MAGI and MHGC achieve ACC scores of 97.23% and 99.48%, respectively, but both obtain a Balance score of only 0.073. In contrast, ACSD achieves a substantially higher Balance score of 0.981. This comparison illustrates that high clustering accuracy alone does not ensure fair representation of sensitive groups within clusters. These methods primarily optimize clustering utility without explicitly constraining sensitive bias.

4.2 t-SNE Visualization

To examine whether the adversarial encoder captures sensitive cluster structure, we visualize the backbone representations $\mathbf{H}$ and the sensitive representations $\mathbf{H}^s$ on the New Bail and Syn-1 datasets using t-SNE (Van der Maaten & Hinton, 2008). As shown in Figure 2, different colors indicate different sensitive groups. In $\mathbf{H}^s$, nodes from the same sensitive group tend to cluster together, suggesting that the adversarial encoder captures grouping patterns associated with sensitive attributes. In contrast, nodes from different sensitive groups are more mixed in $\mathbf{H}$, with less pronounced separation between groups. These observations provide qualitative evidence that ACSD reduces sensitive cluster structure in the learned node representations.

4.3 Ablation Study

To evaluate the contributions of structural adversarial learning and cluster-level fairness regularization, we consider two variants of ACSD: ACSD_v1 removes the sensitive-structure adversary, while ACSD_v2 removes the cluster-level fairness regularization. All remaining components and experimental settings are kept unchanged. As shown in Table 3, removing either component substantially decreases Balance scores, demonstrating their contributions to clustering fairness. Although ACSD_v2 achieves slightly higher ACC on some datasets, this improvement is accompanied by a considerable decline in fairness. These results highlight the importance of structural adversarial

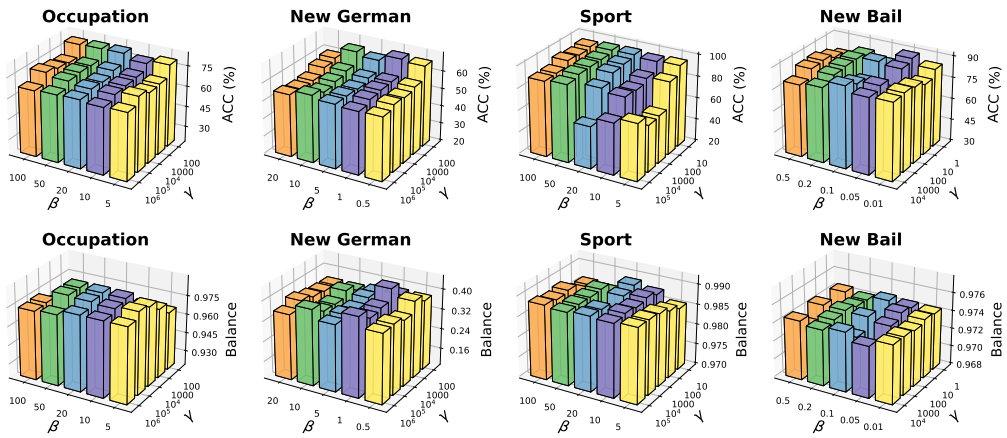

Figure 3: Hyperparameter sensitivity of ACSD to β and γ on four datasets in terms of ACC and Balance. The top row reports ACC, while the bottom row reports Balance.

learning for reducing sensitive bias and suggest that it complements the cluster-level fairness regularizer in promoting fair clustering.

4.4 Analysis of Hyperparameter

We investigate the sensitivity of ACSD to two loss-weight hyperparameters, β and γ, which control the contributions of sensitive-structure adversarial debiasing and cluster-level fairness regularization, respectively. As shown in Figure 3, ACSD maintains consistently high and relatively stable Balance scores across a broad range of β and γ on Occupation, New German, Sport, and New Bail. This indicates that the fairness performance of ACSD is relatively robust to different choices of these hyperparameters. In comparison, ACC exhibits more noticeable variations as β and γ change, suggesting that clustering utility is more sensitive to the weighting of the debiasing and fairness objectives. Overall, these results show that ACSD can maintain stable fairness performance over a relatively broad range of hyperparameter settings, while appropriate parameter selection remains important for preserving clustering utility.

Table 3: Ablation study of different ACSD variants on six benchmark datasets.

Dataset	Metric	ACSD Variants		
		ACSD_v1	ACSD_v2	ACSD
New Bail	ACC	63.09	**70.05**	69.34
	NMI	2.37	9.68	8.41
	BAL	0.845	0.718	**0.974**
New German	ACC	66.85	**68.80**	68.30
	NMI	1.32	0.93	1.89
	BAL	0.336	0.354	**0.435**
New Credit	ACC	62.69	65.71	**66.51**
	NMI	1.01	3.24	1.69
	BAL	0.052	0.072	**0.093**
Occupation	ACC	64.34	**69.19**	67.61
	NMI	1.76	8.64	9.56
	BAL	0.952	0.870	**0.985**
Sport	ACC	53.56	59.43	**59.66**
	NMI	7.04	3.60	4.82
	BAL	0.866	0.901	**0.981**
Syn-1	ACC	78.98	81.74	**82.36**
	NMI	23.63	18.91	32.78
	BAL	0.752	0.908	**0.969**

5 Conclusion

In this paper, we investigate sensitive clustering structure in deep fair graph clustering and propose ACSD, a framework that debiases node representations through structural adversarial learning. Specifically, a sensitive-structure encoder identifies sensitive community structure and pairwise grouping relations through modularity maximization and relation reconstruction, while the clustering backbone opposes this joint objective to reduce sensitive structural bias. Experiments on six benchmark datasets demonstrate the effectiveness of ACSD in improving clustering fairness, while ablation studies support the complementary roles of structural adversarial learning and fairness regularization. In future work, we will investigate the relationship between clustering quality and fairness, motivated by the observation that ACSD achieves lower clustering accuracy than conventional deep graph clustering methods on some datasets.

AI USE STATEMENT

No generative AI tools were used in preparing this manuscript.

REFERENCES

Deyu Bo, Xiao Wang, Chuan Shi, Meiqi Zhu, Emiao Lu, and Peng Cui. Structural deep clustering network. In *Proceedings of the web conference 2020*, pp. 1400–1410, 2020.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Proceedings of the Advances in neural information processing systems*, pp. 5029–5037, 2017.

Fnu Devvrit, Aditya Sinha, Inderjit Dhillon, and Prateek Jain. S3gc: Scalable self-supervised graph clustering. *Advances in Neural Information Processing Systems*, 35:3248–3261, 2022.

Siamak Ghodsi, Seyed Amjad Seyedi, and Eirini Ntoutsi. Towards cohesion-fairness harmony: Contrastive regularization in individual fair graph clustering. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 284–296, 2024.

Yaowen Hu, Wenxuan Tu, Yue Liu, Xinhang Wan, Junyi Yan, Taichun Zhou, and Xinwang Liu. Scalable attribute-missing graph clustering via neighborhood differentiation. In *Proceedings of the International Conference on Machine Learning*, 2025.

Mudi Jiang, Jiahui Zhou, Lianyu Hu, Xinying Liu, Zengyou He, and Zhikui Chen. Adversarial fair multi-view clustering. *Pattern Recognition*, pp. 113708, 2026.

Jia Li, Yanhao Wang, and Arpit Merchant. Spectral normalized-cut graph partitioning with fairness constraints. In *ECAI 2023: 26th European Conference on Artificial Intelligence, September 30–October 4, 2023, Kraków, Poland–Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, pp. 1389–1397. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2023.

Peizhao Li, Han Zhao, and Hongfu Liu. Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9070–9079, 2020.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, Fuchun Sun, and Kunlun He. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (12):9456–9478, 2024.

Yue Liu, Wenxuan Tu, Sihang Zhou, Xinwang Liu, Linxuan Song, Xihong Yang, and En Zhu. Deep graph clustering via dual correlation reduction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 7603–7611, 2022a.

Yue Liu, Jun Xia, Sihang Zhou, Xihong Yang, Ke Liang, Chenchen Fan, Yan Zhuang, Stan Z Li, Xinwang Liu, and Kunlun He. A survey of deep graph clustering: Taxonomy, challenge, application, and open resource. *arXiv preprint arXiv:2211.12875*, 2022b.

Yunfei Liu, Jintang Li, Yuehe Chen, Ruofan Wu, Ericbk Wang, Jing Zhou, Sheng Tian, Shuheng Shen, Xing Fu, Changhua Meng, et al. Revisiting modularity maximization for graph clustering: A contrastive learning perspective. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1968–1979, 2024.

Mark EJ Newman. Mixing patterns in networks. *Physical review E*, 67(2):026126, 2003.

Otavio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. Fairness in deep learning: A survey on vision and language research. *ACM Computing Surveys*, 57(6):1–40, 2025.

Xiaowei Qian, Zhimeng Guo, Jialiang Li, Haitao Mao, Bingheng Li, Suhang Wang, and Yao Ma. Addressing shortcomings in fair graph learning datasets: Towards a new benchmark. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5602–5612, 2024.

Tao Ren, Haodong Zhang, Yifan Wang, Wei Ju, Chengwu Liu, Fanchun Meng, Siyu Yi, and Xiao Luo. Mhgc: Multi-scale hard sample mining for contrastive deep graph clustering. *Information Processing & Management*, 62(4):104084, 2025.

Fei Tian, Bin Gao, Qing Cui, Enhong Chen, and Tie-Yan Liu. Learning deep representations for graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014.

Anton Tsitsulin, John Palowitch, Bryan Perozzi, and Emmanuel Müller. Graph clustering with graph neural networks. *Journal of Machine Learning Research*, 24(127):1–21, 2023.

Wenxuan Tu, Sihang Zhou, Xinwang Liu, Xifeng Guo, Zhiping Cai, En Zhu, and Jieren Cheng. Deep fusion clustering network. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 9978–9987, 2021.

Wenxuan Tu, Qing Liao, Sihang Zhou, Xin Peng, Chuan Ma, Zhe Liu, Xinwang Liu, Zhiping Cai, and Kunlun He. Rare: Robust masked graph autoencoder. *IEEE Transactions on Knowledge and Data Engineering*, 36(10):5340–5353, 2023.

Wenxuan Tu, Renxiang Guan, Sihang Zhou, Chuan Ma, Xin Peng, Zhiping Cai, Zhe Liu, Jieren Cheng, and Xinwang Liu. Attribute-missing graph clustering network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 15392–15401, 2024.

Wenxuan Tu, Sihang Zhou, Xinwang Liu, Zhiping Cai, Yawei Zhao, Yue Liu, and Kunlun He. Wage: Weight-sharing attribute-missing graph autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5760–5777, 2025.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Qianqian Wang, Haiming Xu, Wei Feng, and Quanxue Gao. Adversarial fair incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10011–10019, 2026.

Shiping Wang, Jinbin Yang, Jie Yao, Yang Bai, and William Zhu. An overview of advanced deep graph node clustering. *IEEE Transactions on Computational Social Systems*, 11(1):1302–1314, 2023.

HaiMing Xu, Qianqian Wang, Boyue Wang, and Quanxue Gao. Deep fair multi-view clustering with attention kan. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5061–5070, 2025.

Xihong Yang, Erxue Min, Ke Liang, Yue Liu, Siwei Wang, Sihang Zhou, Huijun Wu, Xinwang Liu, and En Zhu. Graphlearner: Graph node clustering with fully learnable augmentation. In *Proceedings of the ACM international conference on multimedia*, pp. 5517–5526, 2024.

Pengxin Zeng, Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, and Xi Peng. Deep fair clustering via maximizing and minimizing mutual information: Theory, algorithm and metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23986–23995, 2023.

Haodong Zhang, Xinyue Wang, Tao Ren, Yifan Wang, Siyu Yi, Fanchun Meng, Zeyu Ma, Qingqing Long, and Wei Ju. Fairgc: Fostering individual and group fairness for deep graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28194–28202, 2026.

Peng Zhou, Rongwen Li, Zhaolong Ling, Liang Du, and Xinwang Liu. Fair clustering ensemble with equal cluster capacity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (3):1729–1746, 2024.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *Proceedings of the web conference 2021*, pp. 2069–2080, 2021.

A Related Work

Deep fair clustering aims to learn representations that maintain clustering quality while mitigating bias associated with sensitive attributes (Zeng et al., 2023; Xu et al., 2025). A representative line of research adopts adversarial learning to reduce sensitive information in learned representations or clustering assignments. For example, DFC (Li et al., 2020) combines adversarial prediction of protected-group membership from soft cluster assignments with self-supervised structure preservation for visual data. In multi-view clustering, AFMVC (Jiang et al., 2026) introduces a sensitive-attribute discriminator on fused representations and jointly optimizes adversarial fairness, feature reconstruction, and consensus-guided clustering. AFIMVC (Wang et al., 2026) further incorporates adaptive adversarial disentanglement and cross-view contrastive learning to address fairness under incomplete views. These approaches encourage fair clustering by learning fair representations by concealing sensitive-group membership from an adversary.

In the graph domain, deep clustering methods have achieved considerable success by jointly exploiting node attributes and graph topology to learn node representations for clustering (Tian et al., 2014; Liu et al., 2022a; Wang et al., 2023). To address fairness in graph clustering, FairGC (Zhang et al., 2026) adopts a sensitive-label adversarial formulation similar to that used in non-graph fair clustering. It combines high-order affinity-guided contrastive learning for individual fairness with adversarial sensitive-attribute prediction for group fairness. Although graph relationships are incorporated into representation learning, its adversarial fairness objective follows the same sensitive-label prediction paradigm as the aforementioned adversarial methods. In our ACSD framework, sensitive clustering structures serve as the adversarial objective. Under this scheme, the sensitive encoder is optimized to maximize modularity and reconstruct sensitive same-group relations, while the graph encoder operates in opposition to this joint goal, effectively learning debiased representations.

B Dataset Details

German Credit (German). We model clients as nodes and connect pairs with an edge when they exhibit high similarity in credit-account activity patterns. The task is binary node clustering into high- vs. low-credit risk, using gender as the sensitive attribute.

Recidivism (Bail). Nodes represent defendants released on bail between 1990–2009. Edges connect defendants with similar criminal-history and demographic profiles. The task is to predict bail status, with race as the sensitive attribute.

Credit Defaulter (Credit). Nodes correspond to credit-card users; edges connect users with similar purchasing and repayment behavior. The objective is to predict payment default, using age as the sensitive attribute.

Occupation. A Twitter-based user graph where nodes are users and edges denote follow relations. The task is to predict a user's occupation field (*CS vs. Psychology*), with gender (male vs. female) as the sensitive attribute; node features are mean BERT embeddings of users' tweets.

Sport. A Twitter-based athlete graph where nodes are athlete accounts and edges denote follow relations. The task is to predict the sport (*NBA vs. MLB*), with race (black vs. white) as the sensitive attribute; node features are mean BERT embeddings of athletes' tweets.

Syn-1. A synthetic graph dataset with binary labels and a binary sensitive attribute. Node features are formed by concatenating two Gaussian-derived embeddings (one tied to the label and one to the sensitive attribute), and edges are generated via Bernoulli sampling using group–label-dependent edge-type probabilities, enabling controlled evaluation of accuracy–fairness effects from graph structure.

C Proofs of the Theoretical Results

C.1 Proof of Theorem 1

Proof 1 *Let* $\mathbf{s} = (s_1, \ldots, s_n)$ *be sampled uniformly from all label vectors containing exactly* n_g *nodes in sensitive group* g*, for* $g = 1, \ldots, G$*. Independently of* $\mathbf{s}$*, sample a uniformly random*

permutation π of $\{1, \dots, G\}$. Let $\mathbf{e}_g$ denote the gth standard basis vector in $\mathbb{R}^G$, and define:

$$\mathbf{h}_i = \mathbf{e}_{\pi(s_i)}, \qquad i = 1, \dots, n. \tag{19}$$

The same permutation π is shared by all nodes.

For any node i and any $g, a \in \{1, \dots, G\}$,

$$\Pr(\mathbf{h}_i = \mathbf{e}_a \mid s_i = g) = \Pr(\pi(g) = a) = \frac{1}{G}. \tag{20}$$

Therefore, the conditional distribution of $\mathbf{h}_i$ does not depend on s_i, implying:

$$I(\mathbf{h}_i; s_i) = 0, \qquad \forall i. \tag{21}$$

For any distinct nodes i, j, the shared permutation preserves equality of their sensitive labels. Consequently,

$$\begin{aligned} w_{ij} &= \mathbf{h}_i^\top \mathbf{h}_j \\ &= \mathbf{e}_{\pi(s_i)}^\top \mathbf{e}_{\pi(s_j)} \\ &= \mathbb{I}(s_i = s_j) = r_{ij}^s. \end{aligned} \tag{22}$$

Thus, r_{ij}^s is completely determined by w_{ij}, and $\mathsf{H}(r_{ij}^s \mid w_{ij}) = 0$.

Under the specified distribution of $\mathbf{s}$, the probability of a same-group pair is

$$\rho := \Pr(r_{ij}^s = 1) = \sum_{g=1}^{G} \frac{n_g}{n} \frac{n_g - 1}{n - 1} = \frac{\sum_{g=1}^{G} n_g(n_g - 1)}{n(n-1)}. \tag{23}$$

Since $G \geq 2$ and $n_g \geq 2$ for every group, both same-group and different-group pairs have positive probability, giving $0 < \rho < 1$. Hence, $\mathsf{H}(r_{ij}^s) > 0$, and

$$\begin{aligned} I(w_{ij}; r_{ij}^s) &= \mathsf{H}(r_{ij}^s) - \mathsf{H}(r_{ij}^s \mid w_{ij}) \\ &= \mathsf{H}(r_{ij}^s) > 0. \end{aligned} \tag{24}$$

The construction applies to arbitrary group sizes satisfying the stated assumptions and does not require equal sensitive-group proportions.

C.2 Proof of Corollary 1

Proof 2 *Consider the joint distribution of labels and representations constructed in Theorem 1. Conditional on $\mathbf{s}$, independently sample an edge for each unordered pair $i < j$ according to*

$$A_{ij} \mid \mathbf{s} \sim \begin{cases} \text{Bernoulli}(p_{\text{in}}), & s_i = s_j, \\ \text{Bernoulli}(p_{\text{out}}), & s_i \neq s_j. \end{cases} \tag{25}$$

The edge sampling is independent of π conditional on $\mathbf{s}$. Set $A_{ji} = A_{ij}$ and $A_{ii} = 0$, yielding an undirected graph without self-loops.

By construction,

$$\begin{aligned} \Pr(A_{ij} = 1 \mid r_{ij}^s = 1) = p_{\text{in}}, \\ \Pr(A_{ij} = 1 \mid r_{ij}^s = 0) = p_{\text{out}}. \end{aligned} \tag{26}$$

Adding the adjacency matrix in this way does not change the marginal joint distribution of $\mathbf{s}$ and $\{\mathbf{h}_i\}_{i=1}^n$. Therefore, both conclusions of Theorem 1 remain valid.

To establish the final claim, let ρ be defined in Equation equation 23 and let:

$$\eta = \rho p_{\text{in}} + (1 - \rho) p_{\text{out}}. \tag{27}$$

Denote binary entropy by $h_{\text{b}}(x) = -x \log x - (1 - x) \log(1 - x)$, with $0 \log 0 = 0$. Then,

$$\begin{aligned} I(A_{ij}; r_{ij}^s) &= \mathsf{H}(A_{ij}) - \mathsf{H}(A_{ij} \mid r_{ij}^s) \\ &= h_{\text{b}}(\eta) - \rho h_{\text{b}}(p_{\text{in}}) - (1 - \rho) h_{\text{b}}(p_{\text{out}}) \\ &> 0. \end{aligned} \tag{28}$$

The strict inequality follows from the strict concavity of binary entropy, since $0 < \rho < 1$ and $p_{\text{in}} > p_{\text{out}}$.