LEARNING A CLASS-BALANCED CONSENSUS GRAPH FOR FEDERATED LONG-TAILED GRAPH CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated graph clustering aims to collaboratively learn graph representations across multiple clients without sharing raw graph data and discover latent clusters among unlabeled nodes. Existing efforts typically assume that locally learned prototypes provide reliable clustering signals for global aggregation. However, long-tailed data distributions across clients lead to substantial variation in the number of samples among clusters, with nodes from minority classes often exhibiting high heterophily in local graphs and thus likely to receive biased neighboring information, which may result in underrepresented embeddings and unreliable clustering prototypes. Consequently, the global model may fail to sufficiently preserve representations from minority classes during aggregation, further degrading clustering performance. To address this issue, we propose a novel federated long-tailed graph clustering framework termed FedCBG, where clustering feedback guides class-balanced graph optimization, and the refined topology in turn facilitates better clustering. Specifically, to construct the class-balanced graph, we leverage the uploaded cluster-aware signals to derive reliable global prototypes, which are then used on the server to generate an equal number of nodes per cluster and establish their initial topology. Subsequently, we further develop a graph optimization mechanism that leverages global clustering representations to adaptively strengthen reliable connections and suppress unreliable ones for nodes in minority classes. Extensive empirical results on five benchmarks verify the effectiveness and superiority of FedCBG over existing competitors.

1 INTRODUCTION

Federated graph learning (FGL) aims to learn high-quality graph representations through collaborative modeling across multiple clients while preserving data privacy, and has become a fundamental paradigm for distributed graph data mining (Li et al., 2023; Baek et al., 2023). Recently, FGL has achieved success in various fields, including social network analysis (Xie et al., 2021), recommendation systems (Zhang et al., 2024), and medical data analysis (Peng et al., 2023a). The successful application of these methods often relies on an assumption that graph data across different clients exhibit relatively balanced class distributions (Zhang et al., 2021). Nevertheless, this assumption often does not hold in real-world scenarios, where variations in data collection environments and application settings result in highly imbalanced class distributions across clients (Li et al., 2025). In this context, federated long-tailed graph learning has attracted growing attention in recent years.

However, existing FGL approaches for addressing long-tailed class distributions are predominantly supervised and require explicit labels, which are costly to obtain in practice. Once reliable supervision signals are unavailable, class-imbalanced graphs across clients can further exacerbate consensus shifts, making it more challenging for models to discover latent structural patterns of minority classes. Taking PubMed dataset as an example, Fig. 1 shows that classes with fewer samples tend to exhibit higher node heterophily, making them more susceptible to information from other classes and resulting in suboptimal clustering performance. These observations suggest that class imbalance induces unreliable local structures in the absence of label supervision, compromising the representations of minority classes and further biasing global consensus aggregation. Consequently, an effective federated long-tailed graph clustering (FLTGC) framework is urgently needed.

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.

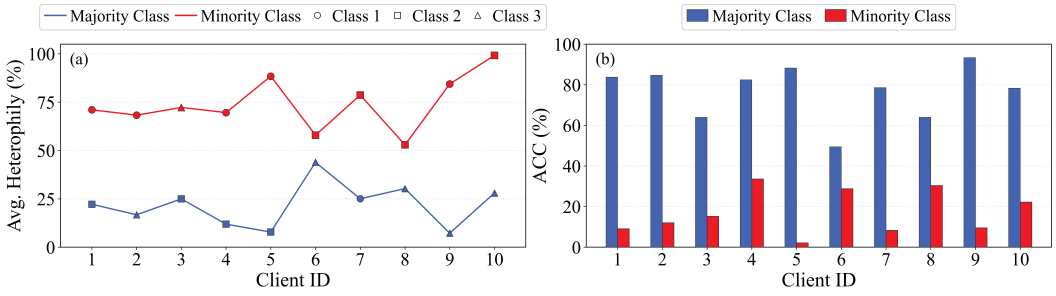

Figure 1: Motivation for federated graph clustering with imbalanced samples. (a) shows the average node heterophily of majority and minority classes across clients on PubMed dataset with a 10-client partition, where minority classes consistently exhibit higher heterophily. (b) reports the clustering accuracy of FedNCN, where minority classes consistently underperform majority classes.

To address the above problem, we revisit FLTGC and propose to adaptively generate more reliable class-balanced samples from the uploaded cluster-aware signals to learn more clustering-friendly graph representations. Nevertheless, achieving this goal remains non-trivial under label absence and privacy protection, mainly posing two key challenges: 1) *how can we ensure the reliability of the uploaded clustering signals without supervised signals from each client, especially in long-tailed graph sample distributions?* and 2) given reliable cluster-level signals on the server, *how can we learn clustering-friendly representations in a balanced manner to alleviate class-distribution imbalance under the FGL setting?* For the first challenge, we introduce an information-theoretic (Liu et al., 2026) reliability assessment mechanism to measure the reliability of uploaded cluster-level information. For the second challenge, inspired by global graph learning (Liu et al., 2025a), we construct a class-balanced graph from reliable cluster-level signals to learn more discriminative global prototypes, alleviating the dominance of majority classes and improving clustering performance.

Based on the above observations, we propose **FedCBG**, a novel federated graph clustering (FGC) framework that adaptively constructs and optimizes a class-balanced graph under clustering guidance in unsupervised distributed settings with long-tailed sample distributions. To achieve this goal, at each client, we first learn node representations with a GNN, and then perform clustering to extract local cluster-aware signals. On the server, we aggregate the uploaded signals into global clustering prototypes, based on which a class-balanced graph is constructed. Subsequently, we design a learnable edge optimizer that evaluates the co-clustering probabilities of node pairs and adaptively refines the corresponding links to improve the structural reliability of minority classes. Finally, the refined class-balanced graph is fed into a GNN to obtain consensus prototypes and model parameters, which are returned to clients for guiding local models to learn more balanced and discriminative clustering representations. Our main contributions are summarized as follows:

- **New task.** To the best of our knowledge, we are the first to investigate FGC on class-imbalanced graphs and reveal that representation bias from cross-client disparities in class distributions is amplified during global aggregation, degrading clustering performance.
- **Novel framework.** A novel FGC framework termed FedCBG is proposed. It not only effectively mitigates bias induced by long-tailed class distributions in unsupervised FGL, but also guides each local model toward better clustering.
- **Superior performance.** Extensive experiments on five benchmark datasets show FedCBG consistently outperforms advanced baselines, validating its superiority and effectiveness.

2 Related Work

2.1 Federated Graph Learning with Imbalanced Samples

FGL with imbalanced samples (Fu et al., 2024; Kong et al., 2024) has recently attracted increasing attention in distributed graph machine learning, as it aims to mitigate global model drift while enabling collaborative graph model training without sharing raw graph data. Specifically, FGPL

(Kong et al., 2024) leverages labeled local data for prototype learning and local graph augmentation, aiming to generate discriminative prototypes that mitigate local class imbalance and domain bias in heterogeneous FGL. Meanwhile, FairFGL (Wu et al., 2025) mitigates fairness bias in FGL by exploiting label signals to transfer reliable cross-client knowledge and improve the representations of minority-class nodes. Additionally, FedSpray (Fu et al., 2024) learns local class-wise structure proxies and aligns them on the server to provide reliable neighboring information for minority-class nodes, which further alleviates biased neighborhood aggregation under heterogeneous label distributions. Previous studies have demonstrated that improving the representations of local minority classes under label supervision can enhance classification performance across clients. In contrast, for FLTGC scenario, we construct a reliable class-balanced graph on the server under the guidance of the clustering objective to mitigate the dominance of majority classes in global consensus learning.

2.2 Attributed Graph Clustering

Attributed graph clustering aims to integrate node attributes and graph structures to group unlabeled nodes into different clusters based on node similarity, and has recently attracted widespread attention in various application domains (Wang et al., 2026; Zhao et al., 2025). For instance, DCLN (Peng et al., 2023b) encourages the node similarity matrix to approximate the high-order adjacency matrix while constraining the feature similarity matrix to approach an identity matrix, which preserves informative neighborhood information and reduces feature redundancy for better clustering. Unlike methods that directly estimate node similarity, HomoCAGC (Chen et al., 2025) exploits homophily to enhance contrastive attributed graph clustering, while exploring and distinguishing the pseudo homophily in graphs. Beyond relation modeling, some methods improve clustering reliability through pseudo-label guidance and data augmentation strategies. ABR (Liu et al., 2025b) first enhances critical features to improve the discriminability of graph representations, and then evaluates the contributions of different representation distributions to refine the clustering target, which in turn generates more reliable pseudo-labels. Despite recent progress, most attributed graph clustering methods adopt a centralized paradigm and assume relatively balanced class distributions, which rarely holds in real-world graphs. However, this assumption is overly strict in real-world graph scenarios. To bridge this gap, we mitigate local class imbalance from a global perspective to enable more balanced and reliable unsupervised graph representation learning under privacy constraints.

3 Methodology

Fig. 2 illustrates the overall architecture of the proposed FedCBG method. The core idea is to construct and optimize a class-balanced consensus graph on the server, aiming to alleviate the adverse effects of long-tailed class distributions across local subgraphs in an FGL setting.

3.1 Notations

Given an undirected graph $\mathcal{G}$ with O clusters, in the FGL setting, $\mathcal{G}$ is partitioned into M disjoint subgraphs distributed across M clients. For a representative client, let $\bar{\mathcal{G}} = \{\bar{\mathcal{V}}, \bar{\mathcal{E}}\}$ denote its local subgraph with $\bar{O}$ clusters, where $\bar{\mathcal{V}}$ and $\bar{\mathcal{E}}$ are the node and edge sets, respectively. Its node attribute matrix and original adjacency matrix are $\bar{\mathbf{X}} \in \mathbb{R}^{\bar{N} \times d}$ and $\bar{\mathbf{A}} \in \mathbb{R}^{\bar{N} \times \bar{N}}$, where $\bar{N}$ is the number of local nodes and d is the attribute dimension. The main notations are summarized in Appendix A.

3.2 Cluster-Aware Signal Extraction

In this part, we design the cluster-aware signal extraction strategy for each local subgraph to preserve the key information of each cluster. First, we employ a GNN to learn node embeddings from $\bar{\mathcal{G}}$ as $\bar{\mathbf{h}}_i^{(l+1)} = \mathrm{UPD}\big(\bar{\mathbf{h}}_i^{(l)}, \mathrm{AGG}(\{\bar{\mathbf{h}}_j^{(l)} \mid j \in \mathcal{V}^i\})\big)$, where $\bar{\mathbf{h}}_i^{(l)} \in \mathbb{R}^{d'}$ denotes an embedding vector of the node i at the l-th layer, $\mathcal{V}^i$ is its neighbor set, and $\mathrm{UPD}(\cdot)$ denotes the node representation update function. After the layer-wise representation propagation, we obtain the final node embeddings $\bar{\mathbf{Z}} \in \mathbb{R}^{\bar{N} \times d''}$. Then, we perform the K-means algorithm to derive the local prototypes $\bar{\mathbf{C}} \in \mathbb{R}^{\bar{O} \times d''}$. Second, to reduce the interference of unreliable prototypes on subsequent global consensus learning, we evaluate the MI-based cluster reliability matrix $\bar{\mathbf{R}} = \mathrm{diag}(\bar{r}^{\bar{\mathbf{c}}_1}, ..., \bar{r}^{\bar{\mathbf{c}}_{\bar{O}}}) \in \mathbb{R}^{\bar{O} \times \bar{O}}$ for $\bar{\mathbf{C}}$. The

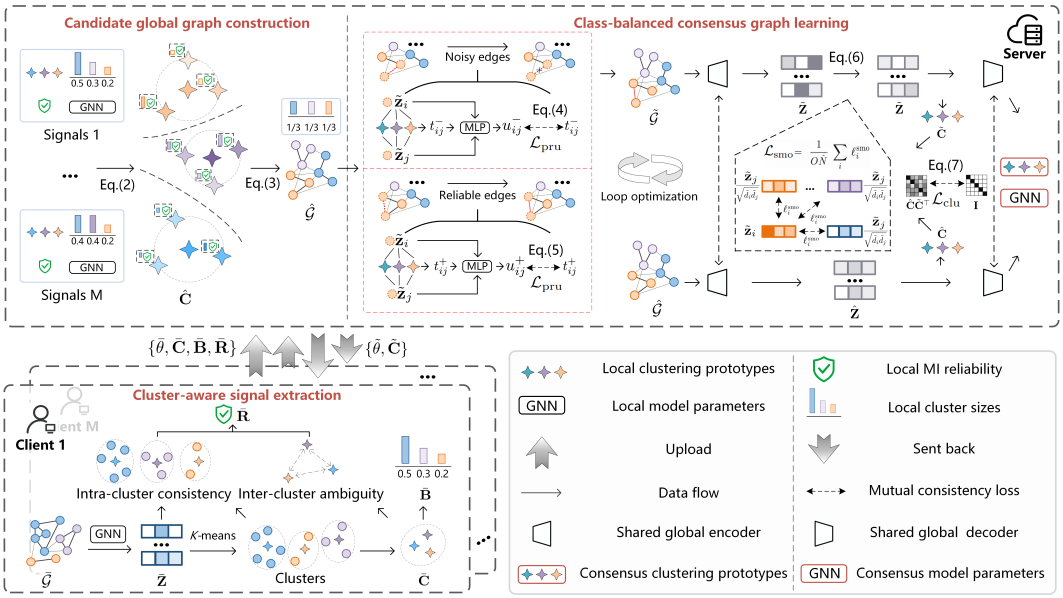

Figure 2: The proposed novel framework consists of three major components, i.e., cluster-aware signal extraction for collecting informative local clustering signals, candidate global graph construction for constructing a class-balanced graph from the extracted signals, and class-balanced consensus graph learning for refining the graph topology and learning clustering-friendly consensus representations. These components are integrated into a unified FGL framework.

reliability of each prototype $\bar{\mathbf{c}}_i \in \bar{\mathbf{C}}$ is computed as

$$\bar{r}^{\bar{\mathbf{c}}_i} = \mathrm{Norm}\left(\frac{\frac{1}{|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_i}|} \sum_{\bar{\mathbf{z}}_j \in \bar{\mathcal{B}}^{\bar{\mathbf{c}}_i}} \frac{\exp\left(\mathrm{Sim}(\bar{\mathbf{z}}_j, \bar{\mathbf{c}}_i)\right)}{\sum_{k=1}^{\bar{O}} \exp\left(\mathrm{Sim}(\bar{\mathbf{z}}_j, \bar{\mathbf{c}}_k)\right)}}{\frac{1}{\bar{O}-1} \sum_{k \neq i}^{\bar{O}} \exp\left(\mathrm{Sim}(\bar{\mathbf{c}}_i, \bar{\mathbf{c}}_k)\right)} \right), \tag{1}$$

where $\bar{\mathbf{z}}_j \in \bar{\mathbf{Z}}$ is the representation of the node j, $\bar{\mathbf{B}} = \mathrm{diag}(|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_1}|, ..., |\bar{\mathcal{B}}^{\bar{\mathbf{c}}_{\bar{O}}}|) \in \mathbb{R}^{\bar{O} \times \bar{O}}$ denotes the cluster size matrix of all prototypes $\bar{\mathbf{C}}$, and $\bar{\mathcal{B}}^{\bar{\mathbf{c}}_i}$ denotes the neighborhood set of the prototype $\bar{\mathbf{c}}_i$. Here, $\mathrm{Norm}(\cdot)$ denotes the normalization function, and $\mathrm{Sim}(\cdot, \cdot)$ denotes the similarity function. Specifically, the intra-cluster score $\frac{1}{|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_i}|} \sum_{\bar{\mathbf{z}}_j \in \bar{\mathcal{B}}^{\bar{\mathbf{c}}_i}} \frac{\exp\left(\mathrm{Sim}(\bar{\mathbf{z}}_j, \bar{\mathbf{c}}_i)\right)}{\sum_{k=1}^{\bar{O}} \exp\left(\mathrm{Sim}(\bar{\mathbf{z}}_j, \bar{\mathbf{c}}_k)\right)}$ measures the compactness between $\bar{\mathbf{c}}_i$ and its assigned nodes, with a larger value indicating greater consistency. In contrast, the inter-cluster score $\frac{1}{\bar{O}-1} \sum_{k \neq i}^{\bar{O}} \exp\left(\mathrm{Sim}(\bar{\mathbf{c}}_i, \bar{\mathbf{c}}_k)\right)$ measures its similarity to other local prototypes, with a smaller value indicating greater separation. Thus, a larger $\bar{r}^{\bar{\mathbf{c}}_i}$ reflects greater prototype reliability in unsupervised settings. Finally, each client uploads the extracted cluster-aware signals to the server, including the local prototypes $\bar{\mathbf{C}}$, the MI-based reliability scores $\bar{\mathbf{R}}$, the sample sizes matrix $\bar{\mathbf{B}}$, and the trained GNN model parameters $\bar{\theta}$, which are used for subsequent class-balanced graph construction and optimization.

In summary, the proposed cluster-aware signal extraction strategy offers two key advantages: 1) it extracts and uploads discriminative clustering knowledge from local subgraphs, providing informative guidance for mitigating class imbalance in distributed scenarios; and 2) it collects the reliability and sample statistics as reliable priors for class-balanced graph construction and optimization.

3.3 Class-Balanced Graph Learning

To alleviate information bias across clients caused by long-tailed class distributions, we propose a global class-balanced graph learning method that integrates cluster-aware knowledge across clients. It comprises candidate global graph construction and class-balanced consensus graph learning.

Candidate Global Graph Construction. To construct the candidate global graph, we first divide the uploaded prototypes $\bar{\mathbf{C}}$ into O regions using K-means, where Ω_i denotes the prototype set in the i-th region. Then, to account for class imbalance across clients, we refine the global prototype $\hat{\mathbf{C}}$ as

$$\hat{\mathbf{c}}_i = \sum_{\bar{\mathbf{c}}_i \in \Omega_i} \frac{\bar{r}^{\bar{\mathbf{c}}_i} |\bar{\mathcal{B}}^{\bar{\mathbf{c}}_i}| \bar{\mathbf{c}}_i}{\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_j} |\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}|}, \tag{2}$$

where $\{\hat{\mathbf{c}}_i\}_{i=1}^{O} \in \hat{\mathbf{C}}$ denotes the global prototype of the i-th region. Theoretical analysis of the weighted prototype aggregation is provided in Appendix B. Next, we construct a class-balanced graph $\hat{\mathcal{G}} = \{\hat{\mathcal{V}}, \hat{\mathcal{E}}\}$ with $O\hat{N}$ samples from the global prototype $\hat{\mathbf{C}}$ in the latent space. Specifically, we sample $\hat{N}$ node embeddings from a Gaussian distribution centered at each prototype, where $\hat{N}$ controls the number of generated nodes per prototype. For each generated embedding $\hat{\mathbf{z}}_i$, we define its neighbor set as $\Gamma_i = \left\{\hat{\mathbf{z}}_j \mid \mathrm{TopK}_{j \in \{1,\ldots,O\hat{N}\},\, j \neq i} \{\mathrm{Sim}(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j)\}\right\}$. After that, we obtain the adjacency matrix $\hat{\mathbf{A}} \in \mathbb{R}^{O\hat{N} \times O\hat{N}}$ among $\hat{\mathbf{z}}_i$ as

$$\hat{\mathbf{Z}} \sim \mathcal{N}\left(\hat{\mathbf{C}}, \hat{\Sigma}\right), \hat{a}_{ij} = \begin{cases} 1, & \text{if } \hat{\mathbf{z}}_j \in \Gamma_i \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where $\hat{\mathbf{Z}} \in \mathbb{R}^{O\hat{N} \times d''}$ denotes the generated node embeddings, $\hat{\Sigma}$ denotes the covariance matrix, and $\hat{a}_{ij} \in \hat{\mathbf{A}}$ indicates the connection between node embeddings $\hat{\mathbf{z}}_i$ and $\hat{\mathbf{z}}_j$. Finally, we decode $\hat{\mathbf{Z}}$ into the original feature space to obtain the reconstructed attribute matrix $\hat{\mathbf{X}} \in \mathbb{R}^{O\hat{N} \times d}$ as $\hat{\mathbf{Z}}^{(l+1)} = \delta(\hat{\mathbf{A}}\hat{\mathbf{Z}}^{(l)}\mathbf{W}_d)$, where $\hat{\mathbf{Z}}^{(l)}$ denotes the node embeddings at the l-th decoder layer, $\mathbf{W}_d \in \mathbb{R}^{d'' \times d''}$ is the decoder weight matrix, and $\delta(\cdot)$ is the activation function. The final decoder layer produces $\hat{\mathbf{X}}$.

Class-Balanced Consensus Graph Learning. Since prototypes of minority classes may be unreliable, the obtained $\hat{\mathbf{A}}$ may inaccurately capture their intra-class structure, introducing irrelevant links and omitting key ones. Thus, we propose a class-balanced consensus graph learning scheme for adaptive structure refinement. First, we initialize $\tilde{\mathcal{G}} = \hat{\mathcal{G}} = \{\hat{\mathcal{V}}, \hat{\mathcal{E}}\}$ and then prune irrelevant edges from $\tilde{\mathcal{G}}$ by estimating their retention probabilities and minimizing the edge-pruning loss $\mathcal{L}_{\mathrm{pru}}$ as

$$\mathcal{L}_{\mathrm{pru}} = \frac{-1}{|\hat{\mathcal{E}}^{\mathrm{pru}}|} \sum_{i,j \in \hat{\mathcal{E}}^{\mathrm{pru}}} \left[\log\left(u_{ij}^{-\,t_{ij}^{-}}\right)(1 - u_{ij}^{-})^{(1-t_{ij}^{-})}\right], \tag{4}$$

where $\hat{\mathcal{E}}^{\mathrm{pru}}$ denotes the candidate edge-pruning set, u_{ij}^{-} is the learnable retention probability of edge (i, j), and t_{ij}^{-} denotes their co-clustering score. They are calculated as $u_{ij}^{-} = \sigma\big(\mathrm{MLP}([\tilde{\mathbf{z}}_i \parallel \tilde{\mathbf{z}}_j \parallel t_{ij}^{-}])\big)$ and $t_{ij}^{-} = \frac{1+\cos(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j)}{2} \sum_{k=1}^{O} \tilde{q}_{ik}\tilde{q}_{jk}$, respectively. Here, $\tilde{\mathbf{z}}_i$ and $\tilde{\mathbf{z}}_j$ denote the representations of reconstructed nodes i and j in $\tilde{\mathcal{G}}$. The $\tilde{q}_{ik}$ and $\tilde{q}_{jk}$ in $\tilde{\mathbf{Q}}$ denote their assignment probabilities to the k-th global prototype. Conversely, to reconstruct the missing key links within minority classes, we estimate the addition probability of each candidate edge and minimize the edge-addition $\mathcal{L}_{\mathrm{add}}$ as

$$\mathcal{L}_{\mathrm{add}} = \frac{-1}{|\hat{\mathcal{E}}^{\mathrm{add}}|} \sum_{i,j \in \hat{\mathcal{E}}^{\mathrm{add}}} \left[\log u_{ij}^{+\,t_{ij}^{+}}(1 - u_{ij}^{+})^{(1-t_{ij}^{+})}\right], \tag{5}$$

where $\hat{\mathcal{E}}^{\mathrm{add}}$ denotes candidate edge-addition set, u_{ij}^{+} is the learnable addition probability of edge (i, j), and t_{ij}^{+} is its co-clustering score. They are computed analogously to u_{ij}^{-} and t_{ij}^{-}. Combining the pruning and addition objectives, we define the edge optimization loss as $\mathcal{L}_{\mathrm{edg}} = \mathcal{L}_{\mathrm{pru}} + \mathcal{L}_{\mathrm{add}}$. Minimizing $\mathcal{L}_{\mathrm{edg}}$ yields the refined graph $\tilde{\mathcal{G}}^{\phi} = \{\hat{\mathcal{V}}, \tilde{\mathcal{E}}\}$, with $\tilde{\mathcal{E}}$ induced by $\tilde{\mathbf{A}} \in \mathbb{R}^{O\hat{N} \times O\hat{N}}$. Second, to preserve local consistency after edge refinement, we introduce a graph smoothness loss $\mathcal{L}_{\mathrm{smo}}$ as

$$\mathcal{L}_{\mathrm{smo}} = \frac{1}{O\hat{N}} \sum_{i=1}^{O\hat{N}} \left\| \tilde{\mathbf{z}}_i - \sum_{j \in \mathcal{V}^i} \frac{1}{\sqrt{\tilde{d}_i \tilde{d}_j}} \tilde{\mathbf{z}}_j \right\|^2, \tag{6}$$

where $\tilde{\mathbf{D}} = \mathrm{diag}(\tilde{d}_1, ..., \tilde{d}_{O\hat{N}}) \in \mathbb{R}^{O\hat{N} \times O\hat{N}}$ denotes the degree matrix of the optimized $\tilde{\mathcal{G}}$. By minimizing $\mathcal{L}_{\mathrm{smo}}$, connected nodes are encouraged to learn locally consistent representations, which

helps reduce unreliable links introduced during graph reconstruction. Third, to promote clustering consistent between $\hat{\mathcal{G}}$ and $\tilde{\mathcal{G}}$, we design a clustering loss $\mathcal{L}_{\text{clu}}$ as

$$\mathcal{L}_{\text{clu}} = \frac{1}{O}\sum_{i=1}^{O}(1 - s_{ii})^2 + \frac{1}{O(O-1)}\sum_{i\neq j} s_{ij}^2, \tag{7}$$

where $\mathbf{S} = \hat{\mathbf{C}}\tilde{\mathbf{C}}^\top \in \mathbb{R}^{O\times O}$ is the correlation matrix, and s_{ii} and s_{ij} denote its diagonal and off-diagonal entries, respectively. Here, $\hat{\mathbf{C}}, \tilde{\mathbf{C}} \in \mathbb{R}^{O\times d''}$ are the global prototypes learned from $\hat{\mathcal{G}}$ and $\tilde{\mathcal{G}}^\phi$. The first term promotes consistency between corresponding prototypes, while the second reduces redundancy among distinct prototypes. Finally, the consensus prototypes $\tilde{\mathbf{C}}$ are learned by minimizing the overall objective $\mathcal{L} = \mathcal{L}_{\text{edg}} + \mathcal{L}_{\text{smo}} + \mathcal{L}_{\text{clu}}$.

The merits of the proposed class-balanced graph learning scheme can be summarized as follows: 1) it explicitly compensates for minority classes, enabling the model to effectively alleviate the dominance bias toward majority classes caused by imbalanced class sample distributions; and 2) it optimizes the class-balanced graph by preserving reliable connections within minority classes while removing unreliable ones, thus promoting more reliable topological structures for better clustering.

3.4 Local-Global Collaborative Learning

After cluster-balanced graph learning, the server returns the consensus prototypes $\tilde{\mathbf{C}}$ and the global model parameters $\tilde{\theta}$ to each client to guide the next round of local learning. Specifically, $\tilde{\mathbf{C}}$ is utilized to update the local cluster center, while $\tilde{\theta}$ is employed to initialize the GNN model parameters of each client. Such local-global knowledge interaction further improves local clustering performance by enhancing the quality of node representations for minority classes in distributed settings. The detailed algorithm for the FedCBG learning process are provided in Appendix C.

4 Experiments

4.1 Experimental Setup

Benchmark Datasets. Following the experimental settings from FedNCN (Liu et al., 2025a), we simulate distributed scenarios by dividing the dataset into 5, 10, and 20 clients, respectively, where each client holds a local subgraph derived from the original graph. Moreover, we adopt five graph datasets as our experimental benchmark datasets, including CiteSeer (Liu et al., 2023), PubMed (Jiang et al., 2024), Amazon-Computer (abbreviated as Am.Co), Amazon-Photo (abbreviated as Am.Ph) (Lin et al., 2021), and Question (Platonov et al., 2024). More details are in Appendix D.1.

Baseline Methods. To demonstrate the superiority and effectiveness of FedCBG, we compare it against four groups of baseline methods. (1) advanced supervised FGL methods, including FedPUB (Baek et al., 2023), FedTAD (Zhu et al., 2024), FedGTA (Li et al., 2023), and FedIIH (Yu et al., 2025); (2) FedNCN (Liu et al., 2025a), the only existing node-level federated graph clustering (NFGC) method; (3) representative federated long-tailed graph classification methods, namely FGPL (Kong et al., 2024), GraphFedMIG (Li et al., 2025), FedSpray (Fu et al., 2024), and FedGM (Zhang et al., 2025); (4) classical federated aggregation strategies, i.e., FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), and FedPer (Arivazhagan et al., 2019). For a fair comparison, the last group utilizes our local model. Further details are presented in Appendix D.2.

Implementation Details. To guarantee a fair comparison, all methods are evaluated under the same hardware environments and experimental configurations. We employ two-layer GNNs on both the client and server for node representation learning, with hidden dimensions of 256 and 32 for the two layers. Model optimization is performed using the Adam optimizer (Xiao et al., 2024) with a learning rate of 3e-3. The local-global communication process consists of 20 communication rounds, where each local model is optimized for 10 epochs in each round. All methods are implemented with PyTorch 2.11.0 and executed on a single NVIDIA GeForce RTX 3090 GPU.

Evaluation Metrics. To comprehensively evaluate the clustering performance of all methods from multiple complementary perspectives, four evaluation metrics are adopted, i.e., ACCuracy (ACC) (Wang et al., 2022), Normalized Mutual Information (NMI) (Liu et al., 2022; Liang et al., 2023), Adjusted Rand Index (ARI) (Yang et al., 2023), and F1 Score (F1) (Tu et al., 2025).

Table 1: Performance comparison across different FGC methods.

Methods	CiteSeer (5 Clients)				CiteSeer (10 Clients)				CiteSeer (20 Clients)			
	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1
FedPUB*	16.8±3.4	0.0±0.0	0.0±0.0	4.1±0.6	9.1±5.2	0.0±0.0	0.0±0.0	3.1±1.6	11.2±6.5	0.0±0.0	0.0±0.0	3.8±2.0
FedTAD*	17.6±3.3	1.2±1.4	0.6±1.0	7.1±2.6	19.1±1.6	4.0±2.6	1.6±1.3	9.9±2.7	17.8±1.5	3.4±3.6	1.1±1.4	6.8±2.4
FedGTA*	24.1±0.8	5.7±1.1	3.5±0.7	20.0±1.1	26.3±0.6	8.4±1.6	4.7±2.0	19.7±0.7	28.1±1.6	12.4±2.1	4.3±2.1	17.1±2.2
FedIIH*	14.2±2.9	0.0±0.0	0.0±0.0	3.7±0.6	17.6±6.1	0.7±0.8	0.4±0.5	4.8±1.1	18.3±7.5	0.0±0.1	-0.1±0.1	5.3±1.9
FedNCN	55.0±1.8	13.3±2.5	19.7±3.7	24.8±1.4	59.0±3.0	17.5±1.3	15.7±4.0	27.1±2.9	57.5±1.6	19.7±1.9	18.0±1.7	32.8±1.1
FGPL*	59.6±2.0	13.7±0.4	9.5±0.7	28.2±1.5	58.1±1.9	13.6±0.8	11.0±2.4	28.9±1.4	61.2±0.7	15.7±0.9	10.4±1.4	31.6±0.9
GraphFedMIG*	32.4±1.3	12.2±1.6	4.0±1.0	24.4±0.8	34.6±0.9	16.3±1.5	5.1±1.4	23.4±1.1	37.1±1.7	24.3±1.8	5.7±1.0	27.6±1.0
FedSpray*	43.7±6.5	20.3±6.3	16.2±6.8	28.2±10.1	35.2±4.9	14.1±4.7	7.2±3.6	26.0±5.2	29.6±1.7	10.8±2.0	3.0±2.1	24.8±3.6
FedGM*	38.0±3.0	16.3±2.9	8.7±3.8	29.9±3.8	35.7±5.9	13.6±4.5	9.8±7.1	28.8±2.8	33.0±2.2	10.0±1.8	7.0±1.5	28.1±2.8
Ours	**64.1±2.3**	**29.7±3.0**	**33.4±5.8**	**32.8±3.3**	**60.4±1.0**	**25.7±1.7**	**23.5±2.4**	**31.9±1.7**	**62.2±0.6**	**27.8±3.0**	**22.4±3.5**	**35.0±0.6**
Methods	**PubMed (5 Clients)**				**PubMed (10 Clients)**				**PubMed (20 Clients)**			
FedPUB*	32.0±9.4	0.0±0.0	0.0±0.0	14.1±3.2	32.0±9.0	0.0±0.0	0.0±0.0	14.0±3.2	28.3±9.3	0.0±0.0	0.0±0.0	12.4±3.3
FedTAD*	34.1±4.7	1.2±1.0	0.8±0.4	26.9±3.5	37.3±4.6	0.4±0.5	0.5±0.9	20.2±3.9	36.5±5.2	0.4±0.7	0.7±1.2	18.2±1.6
FedGTA*	46.2±2.5	6.2±1.9	6.3±1.9	39.6±2.8	46.9±1.7	2.1±0.8	1.8±1.2	33.3±1.9	48.8±1.7	2.5±0.7	2.6±0.8	32.2±1.0
FedIIH*	32.3±9.7	0.0±0.1	0.1±0.3	14.6±3.5	33.6±7.8	0.0±0.0	0.0±0.0	14.6±2.8	28.3±9.3	0.0±0.0	0.0±0.0	12.4±3.3
FedNCN	63.4±1.6	10.5±3.5	13.9±3.6	44.0±2.3	65.0±1.5	10.6±1.8	13.6±4.6	44.4±3.6	66.3±1.7	7.5±0.9	11.0±1.2	41.3±1.3
FGPL*	56.0±2.4	4.5±0.5	4.1±1.1	36.6±0.7	56.4±1.2	3.0±0.7	2.2±1.0	34.7±1.1	57.1±0.7	3.6±0.3	2.5±0.2	34.9±0.9
GraphFedMIG*	37.5±8.7	8.5±1.3	7.3±1.6	39.0±1.7	47.2±1.8	6.8±0.8	5.0±0.8	37.3±1.7	47.2±1.5	6.0±0.9	4.5±0.9	36.1±1.3
FedSpray*	49.1±5.4	11.0±6.1	8.8±5.9	44.4±8.2	43.5±2.9	8.8±3.2	2.8±2.7	36.6±5.6	48.3±1.8	7.1±1.4	4.8±2.5	38.6±2.5
FedGM*	50.1±2.3	7.1±2.8	5.7±2.7	40.1±5.6	42.9±5.2	4.4±2.4	3.3±4.0	39.2±5.0	42.8±3.5	2.6±0.8	2.7±1.2	41.0±2.9
Ours	**70.8±1.8**	**14.3±1.0**	**16.3±1.2**	**46.3±1.7**	**67.6±3.4**	**13.8±2.0**	**16.7±2.0**	**46.8±2.8**	**75.5±1.1**	**11.5±4.9**	**15.9±6.0**	**41.7±4.6**
Methods	**Am.Co (5 Clients)**				**Am.Co (10 Clients)**				**Am.Co (20 Clients)**			
FedPUB*	11.7±13.0	0.0±0.0	0.0±0.0	2.3±2.4	22.4±12.3	0.0±0.0	0.0±0.0	5.0±3.2	14.3±12.1	0.0±0.0	0.0±0.0	4.0±3.5
FedTAD*	9.8±4.4	3.2±1.7	3.3±1.6	3.3±1.5	12.8±12.4	3.7±2.0	4.7±2.6	3.4±1.9	17.8±11.3	5.1±1.3	4.0±0.4	3.8±1.4
FedGTA*	21.4±1.8	12.1±4.0	8.4±3.3	9.0±1.0	21.0±1.5	6.5±1.2	3.7±1.0	6.9±0.5	22.1±1.3	7.3±1.0	3.5±0.8	6.3±0.3
FedIIH*	13.5±7.1	0.0±0.0	0.0±0.0	2.6±1.5	7.8±4.2	0.1±0.1	0.1±0.1	1.9±0.9	7.6±4.6	0.0±0.0	0.1±0.1	2.4±1.8
FedNCN	64.2±1.2	23.5±1.0	21.6±2.3	24.5±2.1	71.2±2.0	21.3±3.1	23.8±4.5	24.0±0.8	70.2±2.6	21.6±1.5	22.9±3.2	32.5±0.8
FGPL*	56.2±4.1	9.7±5.3	9.1±5.8	19.8±0.7	57.8±1.7	10.2±2.2	7.7±2.9	20.2±0.9	59.4±0.9	10.8±1.9	9.3±2.3	22.6±0.6
GraphFedMIG*	14.9±4.8	11.6±2.7	5.4±2.0	14.7±1.4	15.8±5.2	14.6±0.8	6.3±0.7	13.2±0.6	11.1±1.6	11.7±1.4	5.1±1.2	13.1±0.6
FedSpray*	36.9±9.5	28.5±13.0	14.1±15.2	17.8±7.3	35.8±6.3	28.5±11.8	7.6±10.5	21.0±9.5	29.9±4.2	24.9±3.7	2.7±5.0	20.3±3.2
FedGM*	32.0±4.5	20.2±4.0	11.2±4.5	19.1±2.8	31.1±3.2	23.1±2.2	8.7±1.1	22.8±1.2	27.7±1.8	22.9±1.8	8.4±1.5	23.2±1.2
Ours	**66.7±1.1**	**32.7±3.1**	**26.4±3.6**	**28.2±1.2**	**73.0±0.5**	**31.2±1.2**	**32.1±0.9**	**28.0±1.3**	**78.4±0.9**	**26.3±0.7**	**27.2±1.3**	**33.1±0.7**
Methods	**Am.Ph (5 Clients)**				**Am.Ph (10 Clients)**				**Am.Ph (20 Clients)**			
FedPUB*	14.9±7.6	0.0±0.0	0.0±0.0	3.4±1.7	8.4±3.2	0.0±0.0	0.0±0.0	1.9±0.5	12.7±6.8	15.2±0.0	15.2±0.0	6.0±3.6
FedTAD*	13.5±5.0	4.1±6.8	4.0±6.8	3.7±2.5	13.1±3.6	5.4±2.2	5.5±2.5	4.0±1.7	11.8±5.0	8.3±1.0	8.6±1.2	4.3±1.4
FedGTA*	35.5±2.0	14.3±4.0	11.9±4.1	12.0±1.2	32.7±1.1	11.8±0.8	9.6±1.1	10.8±0.6	29.4±1.2	10.8±0.8	7.8±0.9	9.6±0.3
FedIIH*	14.3±4.2	0.0±0.0	0.0±0.0	3.1±1.0	11.7±7.8	0.0±0.0	0.0±0.0	2.7±1.9	12.1±7.8	15.3±0.1	15.2±0.0	4.1±3.5
FedNCN	74.1±3.0	37.2±5.3	43.1±7.0	29.1±2.8	73.5±1.8	29.2±2.1	31.1±3.4	27.6±1.1	66.1±2.1	19.9±1.6	16.7±2.0	32.5±0.9
FGPL*	60.0±1.5	9.5±1.7	8.8±1.3	24.3±1.0	61.1±1.6	8.8±2.4	7.8±3.4	22.6±1.0	60.6±1.3	8.6±1.1	6.1±1.6	24.3±0.9
GraphFedMIG*	24.3±4.1	21.2±5.9	14.8±6.0	19.2±1.6	20.8±1.6	17.9±1.3	8.8±0.8	16.8±0.9	10.7±1.5	12.5±0.8	5.2±0.9	15.7±0.7
FedSpray*	36.0±4.2	32.9±6.4	16.4±5.2	23.7±4.7	38.3±7.5	32.0±8.0	13.9±8.8	27.6±9.8	34.9±3.3	25.9±3.9	5.0±2.3	28.4±5.6
FedGM*	39.3±3.9	29.3±4.6	16.7±3.2	29.9±4.3	29.4±2.0	16.0±1.9	8.2±1.0	26.5±1.6	29.8±3.5	20.0±2.1	9.5±2.4	28.4±1.8
Ours	**77.5±3.1**	**46.6±1.7**	**52.1±4.9**	**31.6±1.1**	**81.3±2.7**	**32.2±4.2**	**34.9±5.2**	**27.9±1.4**	**84.8±1.2**	**26.1±2.0**	**26.8±2.9**	**38.2±1.3**
Methods	**Question (5 Clients)**				**Question (10 Clients)**				**Question (20 Clients)**			
FedPUB*	59.4±46.0	0.0±0.0	0.0±0.0	30.7±22.7	59.4±46.1	0.0±0.0	0.0±0.0	30.7±22.7	40.6±46.0	0.0±0.0	-0.0±0.1	21.5±22.7
FedTAD*	56.1±27.8	0.2±0.1	-0.8±0.8	35.8±12.4	66.3±31.3	0.1±0.1	-0.3±0.3	38.6±14.8	69.1±34.4	0.1±0.1	-0.3±0.4	38.7±13.0
FedGTA*	88.5±4.5	0.6±0.3	3.2±1.6	50.5±1.8	88.1±3.0	0.4±0.1	1.3±0.4	49.6±0.5	86.7±2.4	0.7±0.7	0.7±2.1	47.8±1.5
FedIIH*	77.4±6.2	0.0±0.0	0.0±0.0	30.7±22.7	78.3±3.4	0.0±0.0	0.0±0.0	30.7±22.7	80.1±2.8	0.4±0.7	0.6±1.3	41.6±2.9
FedNCN	94.6±2.0	1.9±0.9	6.4±2.9	52.9±1.7	91.3±4.7	1.6±0.8	4.6±2.3	51.4±2.2	93.0±5.1	1.1±0.8	2.9±3.3	50.8±2.3
FGPL*	44.1±3.7	0.6±0.0	-0.1±0.1	19.5±2.6	56.4±1.1	1.1±0.1	1.9±0.3	22.5±1.8	57.9±2.2	1.1±0.1	1.9±0.4	23.0±0.6
GraphFedMIG*	69.8±4.5	0.3±0.1	0.3±0.1	43.7±1.3	68.1±3.5	0.5±0.1	0.3±0.4	43.5±0.9	68.5±4.1	0.8±0.2	2.3±0.8	46.8±1.6
FedSpray*	77.3±1.2	0.3±0.0	-1.8±0.2	45.0±0.3	77.7±1.4	0.3±0.1	-1.7±0.2	45.1±0.4	77.0±0.5	0.2±0.0	-1.5±0.0	45.0±0.2
FedGM*	60.7±12.1	0.2±0.1	0.2±0.5	40.5±4.5	64.0±9.1	0.2±0.1	-0.7±0.6	40.8±3.2	67.3±6.5	0.1±0.1	0.4±0.4	43.2±2.1
Ours	**94.7±0.3**	**2.5±0.8**	**8.7±1.1**	**54.5±0.5**	**94.4±0.6**	**2.4±0.4**	**7.0±1.0**	**53.6±0.5**	**95.6±0.3**	**3.4±0.2**	**7.6±0.3**	**53.7±0.2**

* Note that these supervised FGL methods are adapted to the unsupervised scenario.

Table 2: Minority-class clustering performance under class imbalance on PubMed (5 Clients).

Client	Minority Class	$n_{\min}/n_{\max}$	ACC		ΔACC	F1		ΔF1
			FedNCN	FedCBG		FedNCN	FedCBG	
1	2	495/2158	36.16	**76.08**	**+39.92**	30.42	**34.67**	**+4.25**
2	2	96/3431	16.67	**20.83**	**+4.16**	14.55	**19.05**	**+4.50**
3	1	665/2105	41.95	**92.18**	**+50.23**	48.02	**74.94**	**+26.92**
4	1	77/3380	18.18	**46.75**	**+28.57**	23.14	**23.30**	**+0.16**
5	3	537/2671	13.04	**32.59**	**+19.55**	14.74	**30.67**	**+15.93**

4.2 Experimental Results

Performance Comparison on the FLTGC Task. To validate the superiority of FedCBG, which is the first FLTGC framework, we compare it with nine advanced FGL methods on five datasets.

Table 3. Performance of FL strategies on FLTGC.

Datasets	Methods	ACC	NMI	ARI	F1	avg. Δ
CiteSeer (10 Clients)	Local	43.6±1.3	19.9±0.7	12.0±1.5	28.5±0.9	-
	FedAvg	43.5±1.4	20.3±0.4	12.6±0.7	28.8±0.6	+0.3
	FedProx	43.6±1.3	19.7±0.8	12.1±1.5	27.4±0.8	-0.3
	FedPer	43.4±1.2	20.0±0.9	12.1±1.5	28.6±0.9	+0.0
	FedCBG	**60.4±1.0**	**25.7±1.7**	**23.5±2.4**	**31.9±1.7**	**+9.4**
PubMed (10 Clients)	Local	59.5±1.5	8.4±0.8	11.2±0.6	42.5±1.6	-
	FedAvg	60.4±0.5	9.2±1.2	12.1±0.8	42.6±0.4	+0.7
	FedProx	59.8±0.9	8.2±0.5	11.0±0.8	43.0±0.9	+0.1
	FedPer	60.8±1.2	10.0±1.7	13.5±2.5	43.5±1.4	+1.6
	FedCBG	**67.6±3.4**	**13.8±1.0**	**16.7±2.0**	**46.8±2.8**	**+5.8**
Am.Co (10 Clients)	Local	49.6±1.9	28.0±0.9	19.2±1.8	**28.1±1.0**	-
	FedAvg	47.2±1.5	27.8±0.8	18.0±1.2	26.8±0.9	-1.3
	FedProx	48.8±1.2	27.4±0.7	17.7±0.9	27.4±0.4	-0.9
	FedPer	47.7±0.6	28.0±1.2	18.4±1.3	27.7±0.7	-0.8
	FedCBG	**73.0±0.5**	**31.2±1.2**	**32.1±0.9**	28.0±1.3	**+9.9**
Am.Ph (10 Clients)	Local	49.2±1.9	26.3±0.4	16.8±0.9	26.6±0.7	-
	FedAvg	48.5±1.6	26.4±0.4	17.0±0.8	26.3±0.5	-0.2
	FedProx	48.0±1.5	26.4±0.8	16.2±1.5	25.8±0.8	-0.6
	FedPer	47.6±1.2	25.9±0.9	16.6±0.8	25.7±1.0	-0.8
	FedCBG	**81.3±2.7**	**32.2±4.2**	**34.9±5.2**	**27.9±1.4**	**+14.4**
Question (10 Clients)	Local	61.6±2.4	0.2±0.0	-0.5±0.1	39.8±0.8	-
	FedAvg	82.3±1.0	0.4±0.1	1.5±0.2	48.3±0.3	+7.9
	FedProx	89.9±1.7	0.9±0.2	4.0±0.9	51.5±0.8	+11.3
	FedPer	68.3±0.9	0.1±0.0	-0.1±0.1	42.6±0.4	+2.5
	FedCBG	**94.4±0.6**	**2.4±0.4**	**7.0±1.0**	**53.6±0.5**	**+14.1**

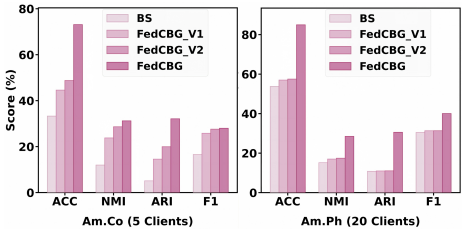

Figure 3: Ablation study of the CGGC strategy and CCGL scheme.

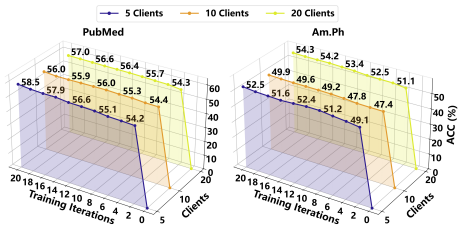

Figure 4: Model convergence.

Notably, for a fair comparison, eight methods originally designed for supervised tasks are adapted into their unsupervised counterparts. Table 1 reports the mean and standard deviation over five runs. From these experimental results, we can draw the following major observations: 1) *Overall performance.* FedCBG achieves the best average performance across all four evaluation metrics on the five datasets. For example, on the Am.Co dataset under 10-client partitions, FedCBG outperforms the second-best method by 1.8%, 2.7%, 8.3%, and 4.0% in terms of ACC, NMI, ARI, and F1, respectively. These results provide strong evidence of the superiority of FedCBG in alleviating sample imbalance in FGC. 2) *Compared with supervised FGL methods.* FedCBG substantially outperforms FedPUB, FedTAD, FedGTA, and FedIIH, whose classification-oriented designs are less suited to unlabeled distributed scenarios. 3) *Compared with NFGC.* FedCBG consistently outperforms FedNCN. This gain can be attributed to the class-balanced graph constructed by FedCBG, which yields more balanced global prototypes to guide local clustering, whereas FedNCN does not explicitly account for minority classes. 4) *Compared with the federated long-tailed graph classification method.* FedCBG also surpasses FGPL, GraphFedMIG, FedSpray, and FedGM across all four metrics on the five datasets. These results indicate that FedCBG effectively mitigates bias from imbalanced class distributions across clients and learns clustering-friendly representations in a balanced manner. Notably, on Question, extreme class imbalance allows majority-dominated ACC to remain high despite insufficient minority-class recovery, whereas NMI and ARI more sensitively reflect overall partition quality. Moreover, a fairer comparison is provided in Appendix E.1.

Minority-Class Performance under Class Imbalance. To evaluate the minority-class clustering performance of FedCBG under class imbalance, we compare FedCBG with FedNCN on PubMed (5 clients) and Question (5 clients), reporting ACC, F1, and the smallest and largest class sizes on each client. The results on PubMed are presented in Table 2, while those on Question are provided in Appendix E.2. As shown, we observe that: 1) all clients exhibit substantial class imbalance, with the majority-to-minority ratio reaching approximately 44:1 on Client 4; and 2) FedCBG consistently outperforms FedNCN on the minority class across all five clients, with the largest gains of 50.23% and 26.92% in ACC and F1, respectively, both on Client 3. These results show that FedCBG mitigates imbalance-induced minority-class representation bias and improves minority-class clustering.

Comparison of Classical FL Aggregation Strategies. To validate the effectiveness of the proposed global aggregation strategy, we compare FedCBG with FedAvg, FedProx, and FedPer. In our experiments, all methods adopt our proposed local model, and the model parameters from each client are uploaded for global sharing. Then, the server performs corresponding global aggregations. Table 3 reports four clustering metrics and the average gains over the local model. From these results, we can observe that FedCBG consistently outperforms the classical strategies, achieving average clustering performance gains of 9.4%, 5.8%, 9.9%, 14.4%, and 14.1% on CiteSeer, PubMed, Am.Co, Am.Ph, and Question, respectively. These results show that the proposed strategy generates

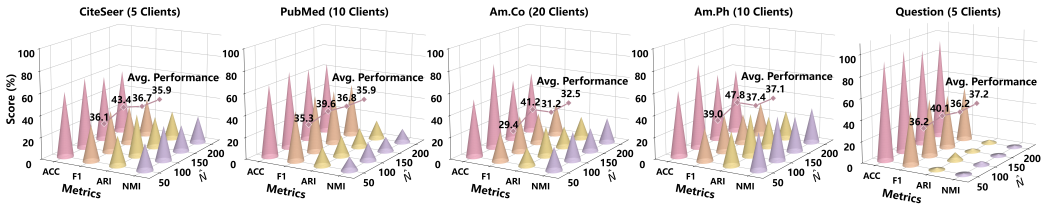

Figure 5: Effect on $\hat{N}$ on the performance of FedCBG across the five datasets.

additional samples for underrepresented classes and learns robust consensus prototypes to improve clustering. Results for the 5-client and 20-client settings are provided in Appendix E.3.

Influence of CGGC Strategy and CCGL Scheme. To evaluate the candidate global graph construction (CGGC) strategy and class-balanced consensus graph learning (CCGL) scheme, we compare FedCBG with its variants on two graph datasets. Concretely, "BS" denotes the local model of FedCBG. "FedCBG_v1" denotes without both components, while "FedCBG_v2" denotes the FedCBG variant without the CCGL scheme. As seen in Fig. 3, we can observe that: 1) *Effectiveness of individual components.* Compared with "BS", "FedCBG_v1" and "FedCBG_v2" improve NMI by 11.80% and 16.67% on Am.Co (5 clients), and by 1.81% and 2.26% on Am.Ph (20 clients), respectively. The consistent superiority of "FedCBG_v2" over "FedCBG_v1" validates CGGC, which generates additional samples for minority classes and alleviates the dominance of majority classes. FedCBG further outperforms "FedCBG_v2", validating CCGL, which refines the generated graph by removing unreliable edges and adding reliable ones; 2) *Effectiveness of component integration.* FedCBG consistently achieves the best performance, confirming the complementary roles of CGGC and CCGL in mitigating class imbalance and learning clustering-friendly prototypes.

Convergence Analysis. To examine the convergence behavior of FedCBG, we record its ACC over training iterations on the PubMed and Am.Ph datasets, as shown in Fig. 4. As seen, the following observations can be concluded: 1) *Convergence trends.* The clustering performance of FedCBG generally improves rapidly during the initial training stage and then gradually converges to a stable level with minor fluctuations. For instance, on the PubMed dataset (5 Clients), the ACC value starts at approximately 54.2% and gradually increases, finally converging to a stable value of around 58.5% after several iterations. 2) *Convergence stability.* Across the evaluated datasets, the ACC of FedCBG stabilizes within 20 training iterations, indicating stable empirical convergence.

Effect of Hyper-Parameter $\hat{N}$. To further analyze the sensitivity of FedCBG to $\hat{N}$, we investigate the clustering performance of FedCBG by varying $\hat{N}$ from 50 to 200 in increments of 50. As shown in Fig. 5, we can find that: 1) *Clustering performance trends with varying $\hat{N}$.* On most datasets, the clustering performance of FedCBG exhibits an increasing-then-decreasing trend as $\hat{N}$ grows. This may be because a relatively small $\hat{N}$ generates insufficient nodes to compensate for sample size disparities among clusters, whereas a relatively large $\hat{N}$ may introduce redundant information, degrading the quality of learned representations; and 2) *Dataset-specific optimal $\hat{N}$ values.* Although the optimal $\hat{N}$ varies across different datasets, it generally around 100. This finding indicates that selecting an appropriate $\hat{N}$ is crucial for FedCBG to construct the class-balanced graph.

5 Conclusion

In this paper, we are the first to explore the problem of FGC under long-tailed sample distributions. We observe that existing studies typically aggregate locally learned centroids, which may be unreliable for minority classes in unlabeled settings and thus bias the global consensus toward majority classes. To address this issue, we propose FedCBG that mines reliable local clustering signals to construct and refine a class-balanced graph at the server, then learns consensus prototypes to guide local models. Extensive experiments on five benchmark datasets demonstrate the effectiveness and superiority of FedCBG in imbalanced FGC tasks. In future work, we plan to improve the generalizability of FedCBG by extending it to more challenging FGC scenarios in which long-tailed distributions coexist with missing attributes.

AI USE STATEMENT

No generative AI tools were employed in the preparation of this manuscript.

REFERENCES

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Jinheon Baek, Wonyong Jeong, Jiongdao Jin, Jaehong Yoon, and Sung Ju Hwang. Personalized subgraph federated learning. In *Proceedings of the International Conference on Machine Learning*, pp. 1396–1415, 2023.

Mansheng Chen, Peiyuan Lai, DeZhang Liao, Changdong Wang, and Jianhuang Lai. Homophily induced contrastive attributed graph clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Xingbo Fu, Zihan Chen, Binchi Zhang, Chen Chen, and Jundong Li. Federated graph learning with structure proxy alignment. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 827–838, 2024.

Bo Jiang, Ziyan Zhang, Sheng Ge, Beibei Wang, Xiao Wang, and Jin Tang. Learning graph attentions via replicator dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7720–7727, 2024.

Xiangjie Kong, Haopeng Yuan, Guojiang Shen, Hanlin Zhou, Weiyao Liu, and Yao Yang. Mitigating data imbalance and generating better prototypes in heterogeneous federated graph learning. *Knowledge-Based Systems*, 296:111876, 2024.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Xinrui Li, Qilin Fan, Tianfu Wang, Kaiwen Wei, Ke Yu, and Xu Zhang. Graphfedmig: Tackling class imbalance in federated graph learning via mutual information-guided generation. *arXiv preprint arXiv:2508.10471*, 2025.

Xunkai Li, Zhengyu Wu, Wentao Zhang, Yinlin Zhu, Ronghua Li, and Guoren Wang. Fedgta: Topology-aware averaging for federated graph learning. *Proceedings of the VLDB Endowment*, pp. 41–50, 2023.

Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):226–238, 2023.

Zhiping Lin, Zhao Kang, Lizong Zhang, and Ling Tian. Multi-view attributed graph clustering. *IEEE Transactions on knowledge and data engineering*, 35(2):1872–1880, 2021.

Jingxin Liu, Renda Han, Wenxuan Tu, Haotian Wang, Junlong Wu, and Jieren Cheng. Federated node-level clustering network with cross-subgraph link mending. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 38540–38556, 2025a.

Jingxin Liu, Xiangyan Tang, Renda Han, Wenxuan Tu, and Ruili Wang. Adaptive feature boosting and distribution refinement for graph clustering. *Pattern Recognition*, 171:112309–112320, 2025b.

Jingxin Liu, Wenxuan Tu, Haotian Wang, Renda Han, Haoyi Li, Junlong Wu, and Xiangyan Tang. Causally-aware attribute completion for incomplete federated graph clustering. pp. 23730–23738, 2026.

Suyuan Liu, Siwei Wang, Pei Zhang, Kai Xu, Xinwang Liu, Changwang Zhang, and Feng Gao. Efficient one-pass multi-view subspace clustering with consensus anchors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7576–7584, 2022.

Yue Liu, Xihong Yang, Sihang Zhou, Xinwang Liu, Siwei Wang, Ke Liang, Wenxuan Tu, and Liang Li. Simple contrastive graph clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282, 2017.

Liang Peng, Nan Wang, Nicha C. Dvornek, Xiaofeng Zhu, and Xiaoxiao Li. Fedni: Federated graph learning with network inpainting for population-based disease prediction. *IEEE Transactions on Medical Imaging*, 42(7):2032–2043, 2023a.

Xin Peng, Jieren Cheng, Xiangyan Tang, Jingxin Liu, and Jiahua Wu. Dual contrastive learning network for graph clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 35 (8):10846–10856, 2023b.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of gnns under heterophily: Are we really making progress?, 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*, 2025.

Penglei Wang, Jitao Lu, Danyang Wu, Rong Wang, and Feiping Nie. Enhance before fusion: Multi-view graph clustering with graph trend filter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Siwei Wang, Xinwang Liu, Suyuan Liu, Jiaqi Jin, Wenxuan Tu, Xinzhong Zhu, and En Zhu. Align then fusion: Generalized large-scale multi-view clustering with anchor matching correspondences. *Advances in Neural Information Processing Systems*, 35:5882–5895, 2022.

Zhengyu Wu, Boyang Pang, Xunkai Li, Yinlin Zhu, Daohan Su, Bowen Fan, Ronghua Li, Guoren Wang, and Chenghu Zhou. Towards unbiased federated graph learning: Label and topology perspectives. *arXiv preprint arXiv:2504.09963*, 2025.

Nachuan Xiao, Xiaoyin Hu, Xin Liu, and Kim-Chuan Toh. Adam-family methods for nonsmooth optimization with convergence guarantees. *Journal of Machine Learning Research*, 25(48):1–53, 2024.

Han Xie, Jing Ma, Li Xiong, and Carl Yang. Federated graph classification over non-iid graphs. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 18839–18852, 2021.

Xihong Yang, Yue Liu, Sihang Zhou, Siwei Wang, Wenxuan Tu, Qun Zheng, Xinwang Liu, Liming Fang, and En Zhu. Cluster-guided contrastive graph clustering network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10834–10842, 2023.

Wentao Yu, Shuo Chen, Yongxin Tong, Tianlong Gu, and Chen Gong. Modeling inter-intra heterogeneity for graph federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 22236–22244, 2025.

Chunxu Zhang, Guodong Long, Tianyi Zhou, Zijian Zhang, Peng Yan, and Bo Yang. Gpfedrec: Graph-guided personalization for federated recommendation. In *Proceedings of the ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4131–4142, 2024.

Hao Zhang, Xunkai Li, Yinlin Zhu, and Lianglin Hu. Rethinking federated graph learning: a data condensation perspective. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 6967–6975, 2025.

Ke Zhang, Carl Yang, Xiaoxiao Li, Lichao Sun, and Siu Ming Yiu. Subgraph federated learning with missing neighbor generation. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 6671–6682, 2021.

Bowen Zhao, Qianqian Wang, Zhengming Ding, and Quanxue Gao. Attribute-missing multi-view graph clustering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25832–25841, 2025.

Yinlin Zhu, Xunkai Li, Zhengyu Wu, Di Wu, Miao Hu, and Rong-Hua Li. Fedtad: Topology-aware data-free knowledge distillation for subgraph federated learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 5716–5724, 2024.

A NOTATIONS

In this section, we provide all notations of the proposed FedCBG in Table 4.

B THEORETICAL JUSTIFICATION

In this section, we theoretically analyze the global prototypes $\hat{\mathbf{C}}$ used to construct the class-balanced graph in Eq. (3). Since these prototypes are estimated in Eq. (2) by jointly considering the MI-based reliability $\bar{\mathbf{R}}$ and cluster sample sizes $\bar{\mathbf{B}}$ of the uploaded local prototypes $\bar{\mathbf{C}}$, we examine the proposed aggregation strategy from two perspectives: relative weight allocation and prototype estimation error. We first analyze how these two factors affect the contributions of prototypes associated with majority and minority classes, and then establish the statistical reliability of the resulting global prototypes.

For each local prototype $\bar{\mathbf{c}}_j \in \Omega_i$, its normalized aggregation weight in Eq. (2) is given by

$$\alpha_{i,j} = \frac{\bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}{\sum_{\bar{\mathbf{c}}_l \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_l} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right|},$$

where $\bar{r}^{\bar{\mathbf{c}}_j}$ denotes the MI-based reliability of $\bar{\mathbf{c}}_j$, and $\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|$ denotes the sample size of its corresponding local cluster. For comparison, we define the weight assigned by count-based aggregation as

$$\beta_{i,j} = \frac{\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}{\sum_{\bar{\mathbf{c}}_l \in \Omega_i} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right|}.$$

Proposition 1. Let $\bar{\mathbf{c}}_j, \bar{\mathbf{c}}_l \in \Omega_i$ be two local prototypes assigned to the same global prototype region, where $\bar{\mathbf{c}}_j$ and $\bar{\mathbf{c}}_l$ are associated with a majority class and a minority class, respectively, satisfying $\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right| > \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right|$. Suppose that $\bar{r}^{\bar{\mathbf{c}}_j} > 0$ and $\bar{r}^{\bar{\mathbf{c}}_l} > 0$. If $\bar{r}^{\bar{\mathbf{c}}_j} \leq \bar{r}^{\bar{\mathbf{c}}_l}$, then

$$\frac{\alpha_{i,j}}{\alpha_{i,l}} \leq \frac{\beta_{i,j}}{\beta_{i,l}}.$$

The inequality is strict when $\bar{r}^{\bar{\mathbf{c}}_j} < \bar{r}^{\bar{\mathbf{c}}_l}$. Therefore, when the prototype associated with the majority class is not more reliable than that associated with the minority class, the proposed aggregation does not increase its relative dominance over the minority-class prototype.

Proof. For count-based aggregation, we have

$$\frac{\beta_{i,j}}{\beta_{i,l}} = \frac{\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}{\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right|}.$$

For the proposed aggregation, we have

$$\begin{aligned}\frac{\alpha_{i,j}}{\alpha_{i,l}} &= \frac{\bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}{\bar{r}^{\bar{\mathbf{c}}_l} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right|} \\ &= \frac{\bar{r}^{\bar{\mathbf{c}}_j}}{\bar{r}^{\bar{\mathbf{c}}_l}} \frac{\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}{\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right|}.\end{aligned}$$

Table 4: Basic notations for the proposed FedCBG.

Notations	Meaning
$\bar{\mathbf{X}} \in \mathbb{R}^{\bar{N} \times d}$	Attribute matrix of each subgraph
$\bar{\mathbf{A}} \in \mathbb{R}^{\bar{N} \times \bar{N}}$	Original adjacency matrix of each subgraph
$\bar{\mathbf{Z}} \in \mathbb{R}^{\bar{N} \times d''}$	Final embeddings matrix of each subgraph
$\bar{\mathbf{C}} \in \mathbb{R}^{\bar{O} \times d''}$	Prototypes of each subgraph
$\bar{\mathbf{B}} \in \mathbb{R}^{\bar{O} \times \bar{O}}$	Cluster size matrix of each subgraph
$\bar{\mathbf{R}} \in \mathbb{R}^{\bar{O} \times \bar{O}}$	Cluster reliability matrix of each subgraph
$\hat{\mathbf{C}} \in \mathbb{R}^{O \times d''}$	Global prototypes
$\hat{\mathbf{Z}} \in \mathbb{R}^{O\hat{N} \times d''}$	Global reconstructed node embeddings matrix
$\hat{\mathbf{A}} \in \mathbb{R}^{O\hat{N} \times O\hat{N}}$	Global reconstructed adjacency matrix
$\mathbf{W}_d \in \mathbb{R}^{d'' \times d''}$	Weight matrix of global decoder
$\hat{\mathbf{X}} \in \mathbb{R}^{O\hat{N} \times d}$	Global reconstructed attribute matrix
$\tilde{\mathbf{Q}} \in \mathbb{R}^{O\hat{N} \times O}$	Soft assignment matrix of the refined class-balanced graph
$\tilde{\mathbf{A}} \in \mathbb{R}^{O\hat{N} \times O\hat{N}}$	Adjacency matrix of the refined class-balanced graph
$\tilde{\mathbf{D}} \in \mathbb{R}^{O\hat{N} \times O\hat{N}}$	Degree matrix of the refined class-balanced graph
$\tilde{\mathbf{C}} \in \mathbb{R}^{O \times d''}$	Consensus prototypes of the refined class-balanced graph
$\mathbf{S} \in \mathbb{R}^{O \times O}$	Correlation matrix between $\hat{\mathbf{C}}$ and $\tilde{\mathbf{C}}$

Since $\bar{r}^{\bar{\mathbf{c}}_j} \leq \bar{r}^{\bar{\mathbf{c}}_l}$, we have

$$\frac{\bar{r}^{\bar{\mathbf{c}}_j}}{\bar{r}^{\bar{\mathbf{c}}_l}} \leq 1.$$

Therefore,

$$\frac{\alpha_{i,j}}{\alpha_{i,l}} \leq \frac{\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}{\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right|} = \frac{\beta_{i,j}}{\beta_{i,l}}.$$

When $\bar{r}^{\bar{\mathbf{c}}_j} < \bar{r}^{\bar{\mathbf{c}}_l}$, we have

$$\frac{\bar{r}^{\bar{\mathbf{c}}_j}}{\bar{r}^{\bar{\mathbf{c}}_l}} < 1,$$

which further gives

$$\frac{\alpha_{i,j}}{\alpha_{i,l}} < \frac{\beta_{i,j}}{\beta_{i,l}}.$$

This completes the proof. □

-

Proposition 2. Under the same setting as Proposition 1, the prototype $\bar{\mathbf{c}}_l$ associated with the minority class receives a larger aggregation weight than $\bar{\mathbf{c}}_j$ associated with the majority class, i.e.,

$$\alpha_{i,l} > \alpha_{i,j},$$

if and only if

$$\frac{\bar{r}^{\bar{\mathbf{c}}_l}}{\bar{r}^{\bar{\mathbf{c}}_j}} > \frac{\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}{\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right|}.$$

Therefore, a minority-class prototype can receive a larger aggregation weight when its advantage in MI-based reliability exceeds its disadvantage in cluster sample size.

Proof. Since $\bar{\mathbf{c}}_j$ and $\bar{\mathbf{c}}_l$ belong to the same global prototype region Ω_i, their aggregation weights share the same normalization denominator. Therefore,

$$\alpha_{i,l} > \alpha_{i,j}$$

holds if and only if

$$\bar{r}^{\bar{\mathbf{c}}_l} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right| > \bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|.$$

Because $\bar{r}^{\bar{\mathbf{c}}_j}\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right| > 0$, dividing both sides by this term gives

$$\frac{\bar{r}^{\bar{\mathbf{c}}_l}}{\bar{r}^{\bar{\mathbf{c}}_j}} > \frac{\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}{\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right|}.$$

Conversely, multiplying both sides of the above inequality by $\bar{r}^{\bar{\mathbf{c}}_j}\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right| > 0$ gives

$$\bar{r}^{\bar{\mathbf{c}}_l}\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right| > \bar{r}^{\bar{\mathbf{c}}_j}\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|,$$

which implies

$$\alpha_{i,l} > \alpha_{i,j}.$$

This completes the proof. □

We next analyze the statistical reliability of the aggregated global prototypes. Throughout the following analysis, expectations are taken conditional on the uploaded reliability scores and cluster sample sizes, which are treated as fixed quantities.

Assumption 1. For each global prototype region Ω_i, all uploaded local prototypes assigned to this region correspond to an ideal global prototype $\hat{\mathbf{c}}_i^\star$. Specifically, each $\bar{\mathbf{c}}_j \in \Omega_i$ satisfies

$$\bar{\mathbf{c}}_j = \hat{\mathbf{c}}_i^\star + \boldsymbol{\xi}_j, \qquad \mathbb{E}[\boldsymbol{\xi}_j] = 0,$$

where $\boldsymbol{\xi}_j \in \mathbb{R}^{d''}$ denotes the estimation error of the uploaded local prototype caused by sample imbalance and clustering errors.

Assumption 2. For each uploaded local prototype $\bar{\mathbf{c}}_j \in \Omega_i$, we assume that $\bar{r}^{\bar{\mathbf{c}}_j} > 0$ and

$$\mathbb{E}\left[\left\|\boldsymbol{\xi}_j\right\|^2\right] \leq \frac{\sigma_{\max}^2}{\bar{r}^{\bar{\mathbf{c}}_j}\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|},$$

where $\sigma_{\max}^2 > 0$ is a uniform constant controlling the estimation errors of the uploaded local prototypes. In addition, for any two distinct local prototypes $\bar{\mathbf{c}}_j, \bar{\mathbf{c}}_l \in \Omega_i$, their estimation errors satisfy $\mathbb{E}[\langle \boldsymbol{\xi}_j, \boldsymbol{\xi}_l \rangle] \leq 0, j \neq l$.

The first condition indicates that the estimation error bound becomes tighter as either the MI-based reliability or the corresponding cluster sample size increases. The second condition excludes the case in which the estimation errors of all local prototypes deviate synchronously in the same direction.

Proposition 3. For each global prototype region Ω_i, the global prototype in Eq. (2) can be written as

$$\hat{\mathbf{c}}_i = \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \alpha_{i,j} \bar{\mathbf{c}}_j,$$

where

$$\alpha_{i,j} = \frac{\bar{r}^{\bar{\mathbf{c}}_j}\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}{\sum_{\bar{\mathbf{c}}_l \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_l}\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right|}.$$

Under Assumptions 1 and 2, $\hat{\mathbf{c}}_i$ is an unbiased estimator of $\hat{\mathbf{c}}_i^\star$, and its estimation error satisfies

$$\mathbb{E}\left[\left\|\hat{\mathbf{c}}_i - \hat{\mathbf{c}}_i^\star\right\|^2\right] \leq \frac{\sigma_{\max}^2}{\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_j}\left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}.$$

Moreover, $\alpha_{i,j}$ minimizes the derived estimation error bound among all convex linear unbiased prototype estimators. Consequently, the proposed aggregation yields an error bound no larger than that of average aggregation.

Proof. From Assumption 1 and $\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \alpha_{i,j} = 1$, we have

$$\begin{aligned}\hat{\mathbf{c}}_i - \hat{\mathbf{c}}_i^\star &= \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \alpha_{i,j}\left(\bar{\mathbf{c}}_j - \hat{\mathbf{c}}_i^\star\right) \\ &= \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \alpha_{i,j} \boldsymbol{\xi}_j.\end{aligned}$$

It follows from $\mathbb{E}[\boldsymbol{\xi}_j] = 0$ that

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{c}}_i] &= \hat{\mathbf{c}}_i^\star + \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \alpha_{i,j} \mathbb{E}[\boldsymbol{\xi}_j] \\ &= \hat{\mathbf{c}}_i^\star.\end{aligned}$$

Therefore, $\hat{\mathbf{c}}_i$ is an unbiased estimator of $\hat{\mathbf{c}}_i^\star$.

According to Assumption 2, its estimation error satisfies

$$\begin{aligned}\mathbb{E}\Big[\|\hat{\mathbf{c}}_i - \hat{\mathbf{c}}_i^\star\|^2\Big] &= \mathbb{E}\left[\left\|\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \alpha_{i,j} \boldsymbol{\xi}_j\right\|^2\right] \\ &= \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \alpha_{i,j}^2 \mathbb{E}\Big[\|\boldsymbol{\xi}_j\|^2\Big] + 2 \sum_{\substack{\bar{\mathbf{c}}_j, \bar{\mathbf{c}}_l \in \Omega_i \\ j<l}} \alpha_{i,j}\alpha_{i,l} \mathbb{E}[\langle \boldsymbol{\xi}_j, \boldsymbol{\xi}_l \rangle] \\ &\le \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \alpha_{i,j}^2 \mathbb{E}\Big[\|\boldsymbol{\xi}_j\|^2\Big] \\ &\le \sigma_{\max}^2 \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \frac{\alpha_{i,j}^2}{\bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}.\end{aligned}$$

Substituting $\alpha_{i,j}$ into the last term yields

$$\begin{aligned}\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \frac{\alpha_{i,j}^2}{\bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|} &= \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \frac{\bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}{\left(\sum_{\bar{\mathbf{c}}_l \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_l} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_l}\right|\right)^2} \\ &= \frac{1}{\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}.\end{aligned}$$

Combining the above results, we obtain

$$\mathbb{E}\Big[\|\hat{\mathbf{c}}_i - \hat{\mathbf{c}}_i^\star\|^2\Big] \le \frac{\sigma_{\max}^2}{\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}.$$

We next consider an arbitrary convex linear estimator

$$\hat{\mathbf{c}}_i^\phi = \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \gamma_{i,j} \bar{\mathbf{c}}_j, \qquad \gamma_{i,j} \ge 0, \qquad \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \gamma_{i,j} = 1.$$

Following the above derivation, its estimation error is bounded by

$$\sigma_{\max}^2 \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \frac{\gamma_{i,j}^2}{\bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}.$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned}1 &= \left(\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \gamma_{i,j}\right)^2 \\ &= \left(\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \frac{\gamma_{i,j}}{\sqrt{\bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}} \sqrt{\bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}\right)^2 \\ &\le \left(\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \frac{\gamma_{i,j}^2}{\bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|}\right) \left(\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_j} \left|\bar{\mathcal{B}}^{\bar{\mathbf{c}}_j}\right|\right).\end{aligned}$$

It follows that

$$\frac{1}{\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_j} \left| \bar{\mathcal{B}}^{\bar{\mathbf{c}}_j} \right|} \leq \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \frac{\gamma_{i,j}^2}{\bar{r}^{\bar{\mathbf{c}}_j} \left| \bar{\mathcal{B}}^{\bar{\mathbf{c}}_j} \right|}.$$

Equality holds if and only if $\gamma_{i,j} \propto \bar{r}^{\bar{\mathbf{c}}_j} \left| \bar{\mathcal{B}}^{\bar{\mathbf{c}}_j} \right|$, which gives

$$\gamma_{i,j} = \frac{\bar{r}^{\bar{\mathbf{c}}_j} \left| \bar{\mathcal{B}}^{\bar{\mathbf{c}}_j} \right|}{\sum_{\bar{\mathbf{c}}_l \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_l} \left| \bar{\mathcal{B}}^{\bar{\mathbf{c}}_l} \right|} = \alpha_{i,j}.$$

Therefore, the proposed aggregation weights minimize the derived estimation error bound.

For average aggregation, $\gamma_{i,j} = 1/|\Omega_i|$, and the corresponding error bound is

$$\frac{\sigma_{\max}^2}{|\Omega_i|^2} \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \frac{1}{\bar{r}^{\bar{\mathbf{c}}_j} \left| \bar{\mathcal{B}}^{\bar{\mathbf{c}}_j} \right|}.$$

Using the Cauchy–Schwarz inequality again, we have

$$\begin{aligned} |\Omega_i|^2 &= \left(\sum_{\bar{\mathbf{c}}_j \in \Omega_i} 1 \right)^2 \\ &\leq \left(\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_j} \left| \bar{\mathcal{B}}^{\bar{\mathbf{c}}_j} \right| \right) \left(\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \frac{1}{\bar{r}^{\bar{\mathbf{c}}_j} \left| \bar{\mathcal{B}}^{\bar{\mathbf{c}}_j} \right|} \right). \end{aligned}$$

Therefore,

$$\frac{\sigma_{\max}^2}{\sum_{\bar{\mathbf{c}}_j \in \Omega_i} \bar{r}^{\bar{\mathbf{c}}_j} \left| \bar{\mathcal{B}}^{\bar{\mathbf{c}}_j} \right|} \leq \frac{\sigma_{\max}^2}{|\Omega_i|^2} \sum_{\bar{\mathbf{c}}_j \in \Omega_i} \frac{1}{\bar{r}^{\bar{\mathbf{c}}_j} \left| \bar{\mathcal{B}}^{\bar{\mathbf{c}}_j} \right|}.$$

Equality holds if and only if all local prototypes in Ω_i have the same value of $\bar{r}^{\bar{\mathbf{c}}_j} \left| \bar{\mathcal{B}}^{\bar{\mathbf{c}}_j} \right|$. Thus, the proposed aggregation yields an estimation error bound no larger than that of average aggregation.

This completes the proof. □

Propositions 1 and 2 characterize the proposed aggregation strategy from the perspective of relative weight allocation. Proposition 1 shows that MI-based reliability prevents a less reliable prototype associated with a majority class from being excessively favored solely because of its larger sample size. Proposition 2 further establishes the condition under which a reliable prototype associated with a minority class can receive a larger aggregation weight despite its limited sample support. From a statistical perspective, Proposition 3 shows that the resulting global prototype is an unbiased estimator and that the proposed weights minimize the derived estimation error bound among convex linear unbiased estimators. These results jointly justify the proposed aggregation strategy in terms of both balanced weight allocation and prototype estimation reliability.

Note that $\hat{\mathbf{c}}_i$ obtained from Eq. (2) is used as an initial semantic center for constructing the class-balanced graph, rather than as the final consensus prototype $\tilde{\mathbf{c}}_i$. Its bounded estimation error helps reduce the propagation of prototype bias during balanced node generation, while the subsequent class-balanced graph learning further refines the prototypes by correcting residual deviations and structural noise.

C ALGORITHMS

In this section, we present the FedCBG learning procedure, as summarized in Algorithm 1.

Algorithm 1 The learning procedure of FedCBG

Input: Complete graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. For each subgraph: attribute matrix $\bar{\mathbf{X}}$; adjacency matrix $\bar{\mathbf{A}}$; server epochs E_{server}; client epochs E_{client}.
Output: Clustering results $\mathbf{Y}$ at clients.
Initialize the $\bar{\theta}$ of all clients.
for $k = 1$ **to** E_{server} **do**
  `// Cluster-Aware Signal Extraction`
  for $l = 1$ **to** E_{client} **do**
    Obtain $\bar{\mathbf{Z}}$ from $\bar{\mathbf{X}}$ and $\bar{\mathbf{A}}$ by GNN.
    Obtain $\bar{\mathbf{C}}$ and $\bar{\mathbf{B}}$ from $\bar{\mathbf{Z}}$ by K-means.
    Obtain $\bar{\mathbf{R}}$ with $\bar{\mathbf{Z}}$, $\bar{\mathbf{C}}$ and $\bar{\mathbf{B}}$ by Eq. (1).
  end for
  Upload $\bar{\mathbf{C}}$, $\bar{\mathbf{R}}$, $\bar{\mathbf{B}}$ and $\bar{\theta}$ to the server.
  `// Class-Balanced Graph Learning`
  Calculate $\hat{\mathbf{C}}$ from $\bar{\mathbf{C}}$, $\bar{\mathbf{R}}$ and $\bar{\mathbf{B}}$ by Eq. (2).
  Generate $\hat{\mathcal{G}}$ and $\tilde{\mathcal{G}}$ from $\hat{\mathbf{C}}$ by Eq. (3).
  Refine $\tilde{\mathcal{G}}^{\phi}$ and train $\tilde{\theta}$ by Eqs. (4)-(7).
  `// Local-Global Collaborative Learning`
  Send back the $\tilde{\mathbf{C}}$ and $\tilde{\theta}$ to each client.
end for
Obtain clustering results $\mathbf{Y}$ at clients by K-means.
return $\mathbf{Y}$

Table 5: Dataset summary.

Datasets	Nodes	Edges	Dimensions	Clusters	Type
CiteSeer	3,327	9,104	3,703	6	Citation Graph
PubMed	19,717	88,648	500	3	Citation Graph
Am.Co	13,752	491,722	767	10	Co-purchase Graph
Am.Ph	7,650	238,162	745	8	Co-purchase Graph
Question	48,921	15,3540	301	2	Question Graph

D Experimental Settings

D.1 Benchmark Datasets

As shown in Table 5, the effectiveness of FedCBG is investigated on five graph datasets, covering two citation graphs (i.e., CiteSeer and PubMed), two co-purchase graphs (i.e., Am.Co and Am.Ph), and one question graph (i.e., Question). Detailed descriptions of these datasets are provided below.

- **Citation Graphs.** CiteSeer (Liu et al., 2023) and PubMed (Jiang et al., 2024) are two widely used citation graph datasets, in which papers are represented as nodes and citation links between papers as edges. It is worth noting that nodes in some citation graph scenarios may represent entities other than papers, such as authors or institutions.

- **Co-purchase Graphs.** In co-purchase graph, Am.Co and Am.Ph (Lin et al., 2021) are two subsets of the Amazon co-purchase graph, where nodes denote products, edges denote frequent co-purchase relationships between products, and node attributes are constructed from bag-of-words encoded product reviews, and class labels correspond to product categories.

- **Question Graph.** Question (Platonov et al., 2024) is a question-related graph dataset. Nodes represent questions, while edges encode the relationships between them. Node attributes are generally derived from textual features, and class labels correspond to the topics of the questions.

D.2 BASELINES AND OUR METHOD

- **FedAvg** (McMahan et al., 2017): This method is a classic FL aggregation framework for distributed collaborative training. Firstly, each local model is trained and its parameters are uploaded to the server, where they are aggregated via weighted averaging to derive the global model parameters. Finally, the updated global model parameters are sent back to each client for updating its local model.
- **FedPer** (Arivazhagan et al., 2019): This method is a personalized FL approach that decomposes model parameters into shared and personalized components. The shared model parameters are uploaded for global aggregation, while the personalized parameters are retained in each client. In this way, the personalized components can be adapted to the specific data distribution of each client.
- **FedProx** (Li et al., 2020): This method is an FL aggregation strategy that introduces a proximal optimization mechanism to address system and statistical heterogeneity. It incorporates a proximal term into the local objective to constrain the deviation between local and global parameters during client optimization, further reducing the impact of heterogeneous environments on FL training.
- **FedPUB** (Baek et al., 2023): This method is a novel personalized FGL framework. In this approach, the central server generates a pseudo-graph with the assistance of ground-truth labels, which is subsequently distributed to each client for representation learning. The learned functional representations are then uploaded to identify densely connected communities. In addition, it employs client-specific sparse masks to select informative subgraph components during local updates, which enhances the quality of global aggregation.
- **FedTAD** (Zhu et al., 2024): This method is a novel supervised FGL framework that jointly considers discrepancies in nodes and structures on local clients to mitigate structural heterogeneity and label distribution variations. It further optimizes the global model by distilling reliable knowledge from local models and alleviates subgraph heterogeneity to enhance local model effectiveness.
- **FedGTA** (Li et al., 2023): This method is a scalable federated graph classification framework for large-scale graph data. It first incorporates topology-aware neighbor features to strengthen local graph representations, which are then leveraged to perform topology-aware global optimization for better handling of complex graph structures.
- **FedIIH** (Yu et al., 2025): This method adopts a hierarchical variational framework at each client to decouple latent subgraph factors and capture intra-subgraph heterogeneity. On the server, it models inter-subgraph similarities to capture cross-subgraph dependencies, alleviating intra- and inter-subgraph heterogeneity and ultimately improving model robustness.
- **FedNCN** (Liu et al., 2025a): This method proposes a novel federated node-level clustering framework with cross-subgraph links mending. On the client, a projector parameterized by an MLP is developed to extract clustering signals, which are uploaded to the server for subsequent inter-client links mending. On the server, a cross-subgraph link mending strategy is designed to estimate graph similarities, enhance correlated links, and remove redundant links via an improved N-Cut approach. Finally, a GNN-based generator is employed to learn consensus prototypes from the refined graph, which are further leveraged to guide local clustering.
- **FGPL** (Kong et al., 2024): This method addresses long-tailed FGL by generating structure-aware minority nodes on the client to enhance minority-class prototypes, while constructing personalized prototypes on the server through class prototype clustering and contrastive alignment, which mitigates class imbalance and majority-class bias.
- **GraphFedMIG** (Li et al., 2025): This method alleviates long-tailed issues by synthesizing minority-class nodes and preserving prototype proximity via mutual information on the client, while performing similarity-based personalized aggregation with generator-driven correction on the server to reduce aggregation bias and improve rare-class recognition.
- **FedSpray** (Fu et al., 2024): This method reduces minority-class neighborhood bias by training personalized GNNs with unbiased soft-label distillation on the client and learning class-level structural proxies to generate reliable soft targets on the server, ultimately alleviating data heterogeneity and minority-class representation deficiency.

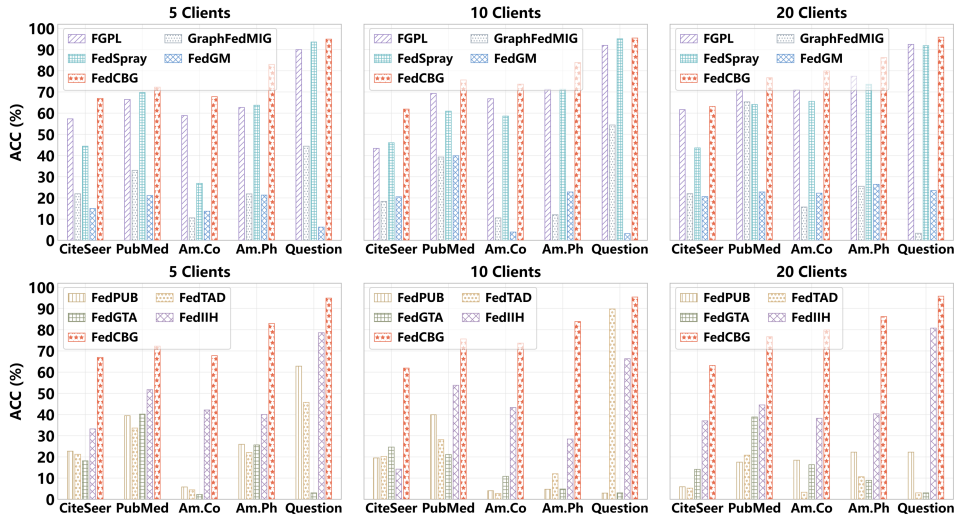

Figure 6: Performance comparison of FedCBG with eight advanced federated graph classification methods on five graph datasets.

Table 6: Minority-class clustering performance under class imbalance on Question (5 Clients).

Client	Minority Class	$n_{\min}/n_{\max}$	ACC		ΔACC	F1		ΔF1
			FedNCN	FedCBG		FedNCN	FedCBG	
1	1	195/9431	64.99	**94.76**	**+29.77**	1.75	**3.45**	**+1.70**
2	1	218/9421	81.05	**96.97**	**+15.92**	4.40	**14.62**	**+10.22**
3	1	284/9216	91.35	**95.44**	**+4.09**	11.80	**19.96**	**+8.16**
4	1	505/9573	89.37	**92.42**	**+3.05**	3.60	**15.86**	**+12.26**
5	1	258/9820	87.05	**93.98**	**+6.93**	3.12	**8.17**	**+5.05**

- **FedGM** (Zhang et al., 2025): This method preserves feature-topology knowledge by condensing real subgraphs via gradient matching on the client and integrating them on the server with class-level gradient optimization, thereby enhancing minority-class representation and global class-level modeling.
- **FedCBG**: The proposed method is a novel federated long-tailed graph clustering framework for graphs with imbalanced class distributions. On the client, local prototypes and cluster-aware information are extracted and uploaded for subsequent class-balanced graph learning. On the server, the uploaded prototypes are aggregated to generate balanced node embeddings and then construct an initial class-balanced graph. Subsequently, a learnable edge optimizer is developed to refine the obtained graph topology, mitigating the dominance of majority classes and facilitating the learning of more discriminative consensus prototypes for improved clustering performance.
- **Local**: This method is a non-FL approach that corresponds to the local model of the proposed FedCBG.

E Experimental Results

E.1 Comparison with Supervised FGL Methods

To further verify the superiority of FedCBG, we compare it with several advanced supervised FGL methods (i.e., FedPUB, FedTAD, FedGTA, FedIIH, FGPL, GraphFedMIG, FedSpray, and FedGM), where each compared method is trained using only 5% labeled data. As shown in Fig. 6, FedCBG consistently achieves the highest ACC across all datasets under different client partitions, outperforming even methods tailored to federated long-tailed graph classification (FGPL, GraphFedMIG, FedSpray, and FedGM). For example, on the Question dataset under the 20-client partition, FedCBG achieves 95.85% ACC, compared with 3.72% for GraphFedMIG, further highlighting its signifi-

Table 7: The performance comparison of different FL strategies in federated long-tailed graph clustering. Specifically, "Local" denotes our local model. In contrast, "FedAvg", "FedProx", and "FedPer" employ different global aggregation strategies, while all clients are equipped with the same local model for a fair comparison.

Datasets	Methods	ACC	NMI	ARI	F1	avg. Δ	Datasets	Methods	ACC	NMI	ARI	F1	avg. Δ
CiteSeer (5 Clients)	Local	47.0±2.4	20.2±0.6	14.0±2.9	28.5±2.0	-	CiteSeer (20 Clients)	Local	49.0±0.2	27.6±0.9	16.9±0.9	34.8±1.0	-
	FedAvg	45.3±1.0	18.6±1.2	14.0±2.0	30.1±0.5	-0.4		FedAvg	49.5±0.8	27.4±1.1	16.9±1.0	34.7±1.2	+0.1
	FedProx	46.1±2.0	19.7±0.8	15.5±1.9	29.7±1.9	+0.3		FedProx	49.1±0.5	27.5±1.1	16.7±1.1	34.7±1.2	-0.1
	FedPer	46.0±2.9	19.6±0.7	12.9±2.5	28.3±2.0	-0.7		FedPer	49.4±0.3	26.7±1.8	16.5±1.6	34.6±0.6	-0.3
	FedCBG	**64.1±2.3**	**29.7±3.0**	**33.4±5.8**	**32.8±3.3**	**+12.6**		**FedCBG**	**62.2±0.6**	**27.8±3.0**	**22.4±3.5**	**35.0±0.6**	**+4.8**
PubMed (5 Clients)	Local	55.2±2.5	9.8±0.9	10.1±1.9	41.3±2.5	-	PubMed (20 Clients)	Local	59.9±0.5	6.6±0.6	7.8±1.3	39.0±0.9	-
	FedAvg	57.6±1.2	11.0±2.6	13.1±2.6	43.4±2.0	+2.2		FedAvg	59.7±0.5	6.9±0.7	8.2±1.4	38.8±0.7	+0.1
	FedProx	55.7±1.6	11.8±3.5	12.6±2.7	43.7±1.1	+1.9		FedProx	59.7±0.4	7.2±0.2	8.8±0.2	39.1±0.4	+0.4
	FedPer	55.8±1.5	10.7±1.9	11.8±1.6	43.3±1.8	+1.3		FedPer	60.2±0.2	7.2±0.3	8.6±0.3	39.6±0.5	+0.6
	FedCBG	**70.8±1.8**	**14.3±1.0**	**16.3±1.2**	**46.3±1.7**	**+7.8**		**FedCBG**	**75.5±1.1**	**11.5±4.9**	**15.9±6.0**	**41.7±4.6**	**+7.8**
Am.Co (5 Clients)	Local	50.0±1.7	33.9±1.0	27.1±1.3	30.9±0.8	-	Am.Co (20 Clients)	Local	48.8±1.0	21.5±0.1	15.5±0.9	29.0±0.5	-
	FedAvg	49.5±0.9	34.0±0.4	26.4±0.3	31.7±0.6	-0.1		FedAvg	48.4±0.3	20.8±0.2	14.5±0.2	28.8±0.1	-0.6
	FedProx	51.4±3.0	**35.3±1.6**	**28.9±2.6**	**32.0±2.7**	+1.4		FedProx	49.0±0.8	20.9±0.5	14.8±0.7	28.3±0.3	-0.5
	FedPer	50.9±1.6	34.2±0.4	26.6±0.1	31.7±1.2	+0.4		FedPer	49.6±0.5	21.7±0.6	15.9±0.6	28.8±0.8	+0.3
	FedCBG	**66.7±1.1**	32.7±3.1	26.4±3.6	28.2±1.2	**+3.0**		**FedCBG**	**78.4±0.9**	**26.3±0.7**	**27.2±1.3**	**33.1±0.7**	**+12.6**
Am.Ph (5 Clients)	Local	55.4±3.1	40.0±3.5	34.2±6.0	27.7±1.3	-	Am.Ph (20 Clients)	Local	49.9±1.5	17.9±0.4	11.1±0.6	27.8±0.6	-
	FedAvg	58.6±3.1	40.5±2.4	37.3±3.8	29.4±1.5	+2.1		FedAvg	48.7±2.0	17.8±0.1	10.7±0.2	27.5±0.7	-0.5
	FedProx	58.6±2.6	41.3±2.9	37.6±5.4	28.5±1.6	+2.2		FedProx	48.7±1.0	17.6±0.4	10.3±0.6	27.7±0.5	-0.6
	FedPer	58.3±1.7	41.7±1.3	38.2±2.5	29.1±0.3	+2.5		FedPer	50.2±0.2	18.0±0.3	10.8±0.3	28.5±0.4	+0.2
	FedCBG	**77.5±3.1**	**46.6±1.7**	**52.1±4.9**	**31.6±1.1**	**+12.6**		**FedCBG**	**84.8±1.2**	**26.1±2.0**	**26.8±2.9**	**38.2±1.3**	**+17.3**
Question (5 Clients)	Local	78.4±3.8	0.3±0.1	-0.9±0.2	44.4±0.8	-	Question (20 Clients)	Local	65.1±0.3	0.4±0.0	-0.3±0.1	41.0±0.1	-
	FedAvg	91.5±1.9	0.1±0.1	0.2±0.6	49.3±0.1	+4.7		FedAvg	87.8±0.2	0.7±0.0	2.0±0.2	49.5±0.2	+8.5
	FedProx	84.3±3.5	0.0±0.0	-0.2±0.5	47.9±0.8	+2.5		FedProx	65.4±1.6	0.5±0.2	0.6±1.2	41.7±1.6	+0.5
	FedPer	87.2±2.3	0.2±0.1	0.3±0.5	48.6±0.7	+3.5		FedPer	70.7±1.5	0.4±0.1	0.2±0.3	43.1±0.5	+2.1
	FedCBG	**94.7±0.3**	**2.5±0.8**	**8.7±1.1**	**54.5±0.5**	**+9.6**		**FedCBG**	**95.6±0.3**	**3.4±0.2**	**7.6±0.3**	**53.7±0.2**	**+13.5**

cant advantage on the FLTGC task. In addition, FedCBG also consistently outperforms FedPUB, FedTAD, FedGTA, and FedIIH, indicating that limited supervision cannot adequately alleviate the representation bias caused by class imbalance. In contrast, FedCBG directly addresses this issue by constructing a class-balanced graph that supports reliable sample generation and balanced consensus prototype learning.

E.2 Minority-Class Performance under Class Imbalance

In the main text, we compare FedCBG with FedNCN in terms of minority-class clustering performance on PubMed (5 clients). To further examine its generalizability, we report the corresponding results on Question (5 clients) in Table 6. For each client, the minority class is defined as the least populated class, and $n_{\min}/n_{\max}$ denotes the sample ratio between the minority and majority classes. As shown, Question exhibits severe client-wise class imbalance, with imbalance ratios ranging from approximately 1:19 to 1:48. Nevertheless, FedCBG consistently outperforms FedNCN in both ACC and F1 across all clients. In particular, FedCBG improves ACC by 29.77% on Client 1 and F1 by 12.26% on Client 4. Although the absolute minority-class F1 scores remain relatively low, reflecting the difficulty of recovering minority-class patterns under extreme class imbalance, the consistent improvements over FedNCN demonstrate that FedCBG effectively mitigates minority-class representation bias across different client-wise class distributions.

E.3 Comparison of FL Aggregation Strategies

In the main text, we validate the superiority of FedCBG over three classical FL aggregation strategies, i.e., FedAvg, FedProx, and FedPer, under the 10-client setting. To further investigate the superiority of FedCBG, we conduct additional experiments under the 5-client and 20-client settings using multiple clustering metrics on five benchmark datasets, with the results summarized in Table 7. Based on these experimental results, we can draw the following observations: 1) under the 5-client setting, FedCBG obtains average performance gains of 12.6%, 7.8%, 3.0%, 12.6%, and 9.6% on the CiteSeer, PubMed, Am.Co, Am.Ph, and Question datasets, respectively; 2) under the 20-client setting, FedCBG still achieves consistent performance improvements of 4.8%, 7.8%, 12.6%, 17.3%, and 13.5%, respectively. These results demonstrate that FedCBG consistently outperforms other FL aggregation strategies employing the same local model under different client partition settings. This superiority can be attributed to the following aspects. First, the proposed cluster-aware signal ex-

traction mechanism effectively captures and uploads reliable clustering signals from local subgraphs, providing informative guidance for subsequent global class-balanced graph learning. Second, the proposed class-balanced graph learning scheme effectively removes incorrectly introduced noisy links while recovering overlooked reliable links associated with minority classes. Finally, the refined graph is fed into the global model to learn consensus prototypes and model parameters, which in turn are transmitted to each client for better clustering.