FEDUGCN: PROBABILISTIC FEDERATED CLUSTERING WITH UNKNOWN GLOBAL CLUSTER NUMBER

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated clustering aims to uncover global clustering structures from distributed unlabeled data without centralizing raw samples. Existing methods typically assume that the global number of clusters is known and recover global clustering structures by aggregating client-specific cluster centers. However, the global cluster number is often unknown in practical settings, while heterogeneous client distributions obscure the global clustering structure and complicate the identification of corresponding clusters across clients. To address this issue, we propose a probabilistic **Fed**erated clustering framework for **U**nknown **G**lobal **C**luster **N**umbers (**FedUGCN**), which jointly infers the number of global clusters and identifies the corresponding global clustering structure through hierarchical aggregation of local clusters across clients and federated probabilistic modeling. Specifically, each client first summarizes its local clusters with cluster-level statistics. The server then employs the Bayesian information criterion to assess evidence for merging clusters across clients and applies constrained hierarchical aggregation to construct candidate global structures. For each candidate, federated probabilistic optimization estimates shared global cluster distributions and client-specific cluster compositions, refining cross-client cluster correspondences while preserving local heterogeneity. Finally, the integrated completed likelihood criterion is designed to select the global number of clusters and latent clustering structure. Extensive experiments demonstrate that FedUGCN achieves competitive clustering performance without requiring prior knowledge of the global cluster number.

1 INTRODUCTION

Federated Learning (FL) has emerged as a privacy-preserving distributed learning paradigm that enables multiple clients to collaboratively optimize a shared global model while keeping their local data decentralized Xue et al. (2025); Yan et al. (2025b); Chen et al. (2026a); Yao et al. (2026); Tang et al. (2026); Zeng et al. (2026). However, the lack of reliable labels in real-world scenarios limits the applicability of conventional supervised federated learning methods. To address this issue, Federated Clustering (FC) integrates unsupervised clustering into distributed learning Ding et al. (2026); He et al. (2026); Yan et al. (2025a); Zhang et al. (2026b); Li et al. (2026), enabling the discovery of consistent global clustering structures from decentralized local structural information without directly sharing raw data. By uncovering global clustering structures from unlabeled data, FC has shown considerable potential for a wide range of applications, including edge intelligence, industrial Internet of Things, and recommender systems Zhang et al. (2026a); Liu et al. (2026); Cheng et al. (2025); Zhang et al. (2026c).

Existing federated clustering methods typically assume that the global number of clusters is predefined and construct global clustering structures by aggregating local cluster centers across clients Dennis et al. (2021); Pan et al. (2023); Wang et al. (2025b). To relax this assumption, recent studies have explored unknown-K federated clustering through adaptive search strategies or local cluster associations Zhang et al. (2025); Ibrahim et al. (2026). However, existing methods still primarily rely on local cluster centers or structural similarities to associate clusters across clients, which inadequately capture the distributional characteristics of heterogeneous local clusters and can lead to unreliable cross-client cluster correspondences. This limitation further complicates global structure inference when the global cluster number is unknown. Consequently, reliably associating hetero-

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.

geneous local clusters across clients while jointly inferring the number of global clusters and the corresponding global clustering structure remains a fundamental challenge in federated clustering.

An intuitive solution is to move beyond local cluster centers and structural similarities by leveraging distributional information to identify cross-client cluster correspondences and infer the global clustering structure. However, achieving this goal poses two key challenges: 1) Heterogeneous client distributions can induce substantial shifts among local cluster distributions, making it difficult to establish reliable cluster correspondences across clients. 2) Without prior knowledge of the global cluster number, cluster number estimation and global structure inference become inherently coupled, making it challenging to recover a coherent global clustering structure from fragmented local clustering information. To address the first challenge, inspired by distributional statistics and mixture modeling in federated learning Zhou et al. (2024); Wu et al. (2023), we characterize local cluster distributions using cluster-level statistics. Despite heterogeneity-induced distribution shifts, local clusters corresponding to the same latent global cluster can retain statistical consistency, enabling reliable cross-client cluster correspondence. To address the second challenge, following the structure inference principle of conventional unknown-K clustering Vardakas et al. (2026); Stirn & Knowles (2025), we jointly infer the global cluster number and corresponding clustering structure through probabilistic optimization over candidate global structures, rather than treating cluster number estimation as an independent search problem.

Motivated by these insights, we propose a probabilistic **Fed**erated clustering framework for an **U**nknown **G**lobal **C**luster **N**umber (**FedUGCN**), which unifies cross-client cluster correspondence discovery and global structure inference under heterogeneous client distributions. To this end, we first characterize local clusters using cluster-level statistics, enabling meaningful comparison across clients at the distribution level rather than relying solely on cluster centers or structural similarities. Next, we evaluate cross-client cluster merge evidence and construct candidate global clustering structures through constrained hierarchical aggregation. Instead of estimating the global cluster number independently, FedUGCN jointly estimates shared global cluster distributions and client-specific cluster compositions over candidate structures through federated probabilistic inference. This process refines cross-client cluster correspondences while preserving local heterogeneity, enabling unified inference of the global cluster number and corresponding global clustering structure. Finally, the candidate structures are compared through model selection to determine the most plausible global clustering structure. The main contributions of this work are summarized as follows:

- **New framework.** We propose FedUGCN, a probabilistic federated clustering framework that addresses unknown-K federated clustering by unifying cross-client cluster correspondence discovery and global structure inference under heterogeneous client distributions.
- **Novel techniques.** We develop a two-stage method consisting of Distribution-Aware Global Structure Discovery to construct candidate structures and Federated Probabilistic Structure Optimization to jointly infer the global cluster number and corresponding global clustering structure.
- **Better performance.** Extensive experiments across multiple benchmark datasets demonstrate that FedUGCN achieves competitive clustering performance against state-of-the-art federated clustering methods without requiring prior knowledge of the global cluster number.

2 Related Work

2.1 Federated Clustering

Federated Clustering (FC) aims to learn global clustering structures from distributed unlabeled data without sharing raw client data. Early methods typically assume a known global cluster number and construct global clusters by aggregating client-side clustering results. For example, k-FED Dennis et al. (2021) performs local k-means clustering and aggregates the obtained centers through a single communication round, while FFCM Stallmann & Wilbik (2022) extends fuzzy c-means to federated scenarios by aggregating local centers with weighted averaging and server-side clustering. Beyond direct center aggregation, NNFC Wang et al. (2024) incorporates privacy constraints and exploits local neighborhood relationships to extract stable anchors for global cluster recovery. To alleviate the reliance on prior knowledge of the global cluster number, recent studies have investigated federated clustering with unknown-K. AFCL Zhang et al. (2025) determines the cluster number via collaborative seed refinement. FedGEM Ibrahim et al. (2026) infers potential clusters using uncertainty

relationships among local mixture components in a mixture model. However, existing methods still establish cross-client cluster associations primarily based on centroids, distances, or parameter-space relationships, which struggle to capture heterogeneous cluster distributions and reliable cross-client correspondences. To address this limitation, we propose FedUGCN, a probabilistic federated clustering framework that jointly infers the global cluster number and latent clustering structures through distribution-aware aggregation and probabilistic modeling.

2.2 BIC-BASED MODEL SELECTION FOR CLUSTERING

Model selection has been widely studied for clustering with an unknown number of clusters. Among existing criteria, the Bayesian Information Criterion (BIC) Stenger et al. (2026); Yi et al. (2026); Neath & Cavanaugh (2012) provides a principled way to compare candidate models by balancing model fit and complexity. Given a statistical model $\mathcal{M}$, BIC is defined as:

$$\mathrm{BIC}(\mathcal{M}) = -2\ell(\widehat{\Theta}_{\mathcal{M}}) + p_{\mathcal{M}} \log N, \tag{1}$$

where $\ell(\widehat{\Theta}_{\mathcal{M}})$ denotes the maximized log-likelihood, $p_{\mathcal{M}}$ is the number of free parameters, and N is the sample size. A lower BIC indicates a preferable model by balancing goodness of fit and model complexity. BIC is widely used in mixture modeling and model-based clustering to determine the number of latent components Wang et al. (2026); Chen et al. (2026b). More recent studies employ model-selection criteria to compare candidate clustering structures under unknown-K settings Zhang & Chen (2025). These studies suggest cluster-number estimation can be integrated with global clustering structure selection rather than treated as an independent search problem.

3 METHOD

We present FedUGCN, a probabilistic federated clustering framework for jointly inferring the unknown global cluster number and latent clustering structure from heterogeneous distributed data. As illustrated in Figure 1, FedUGCN consists of two stages: Distribution-Aware Global Structure Discovery, which constructs a nested sequence of candidate global structures from local cluster-level statistics, and Federated Probabilistic Structure Optimization, which refines each candidate using sample-level probabilistic evidence and selects the final global structure.

3.1 PRELIMINARIES

We consider a federated clustering system with G clients and a central server, where the goal is to infer global clustering structures from distributed unlabeled data while preserving data privacy. Client g privately maintains an unlabeled dataset:

$$\mathcal{D}_g = \left\{ \mathbf{x}_i^{(g)} \right\}_{i=1}^{n_g}, \qquad \mathbf{x}_i^{(g)} \in \mathbb{R}^d, \tag{2}$$

where n_g denotes the number of local samples and d is the feature dimension. Let $N = \sum_{g=1}^{G} n_g$ denote the total number of samples. Each client contains K_g local clusters, whereas the global number of clusters $K^\star$ is unknown to the server. No ground-truth value of $K^\star$ or sample label is involved in global structure inference. Due to heterogeneous and incomplete local observations, the correspondence between independently discovered local clusters is unknown. Our objective is to estimate the global cluster number $\widehat{K}$ and the associated global distribution parameters:

$$\widehat{\Theta} = \left\{ \left(\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k \right) \right\}_{k=1}^{\widehat{K}}, \tag{3}$$

where $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\boldsymbol{\Sigma}}_k$ denote the estimated mean and covariance matrix of the k-th global cluster, respectively, and $\widehat{\Theta}$ denotes the resulting set of global cluster distribution parameters. Each client is allowed to maintain a heterogeneous cluster composition. The unknown global model order and cross-client cluster correspondence are intrinsically coupled: reliable global structure inference requires identifying which local clusters may correspond to the same latent global component, while such correspondences themselves depend on the underlying global structure.

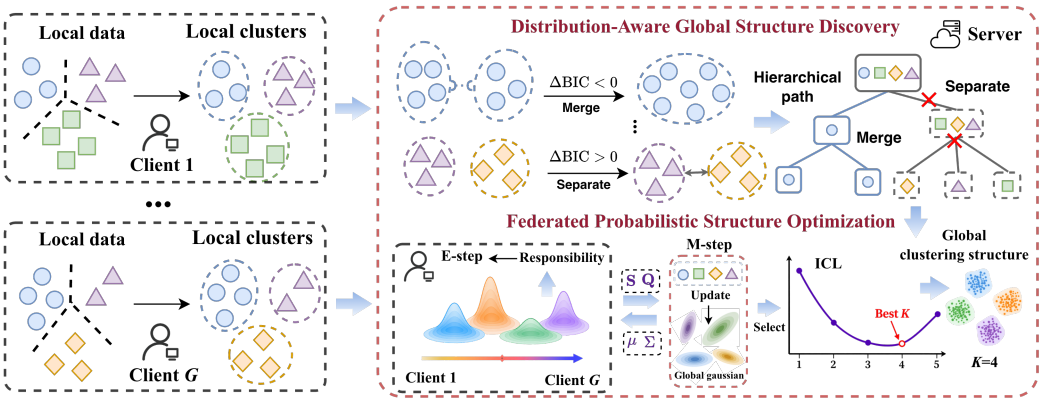

Figure 1: Overview of the proposed FedUGCN framework. Each client performs local clustering and constructs cluster-level statistics to characterize heterogeneous local structures. The server identifies cross-client cluster correspondences through distribution-aware hierarchical aggregation and constructs candidate global structures. Subsequently, federated probabilistic optimization jointly estimates shared global cluster distributions and client-specific cluster compositions for each candidate structure. The final global cluster number and corresponding global clustering structure are selected according to the integrated completed likelihood criterion.

3.2 Distribution-Aware Global Structure Discovery

The first stage constructs a sequence of candidate global clustering structures without centralizing raw client data. Instead of representing a local cluster solely by its centroid, FedUGCN characterizes its empirical distribution using compact cluster-level statistics and evaluates cross-client associations through distributional model comparison. Client g partitions its local data into $\{\mathcal{C}_c^{(g)}\}_{c=1}^{K_g}$ using K-means. To capture distributional information beyond centroid location, each local cluster $\mathcal{C}_c^{(g)}$ is summarized by $\mathcal{S}_c^{(g)} = (n_c^{(g)}, \boldsymbol{\mu}_c^{(g)}, \mathbf{v}_c^{(g)})$, where $n_c^{(g)}$, $\boldsymbol{\mu}_c^{(g)}$, and $\mathbf{v}_c^{(g)}$ denote its size, empirical mean, and element-wise variance, respectively. This representation captures both location and dispersion information while remaining closed under hierarchical aggregation. For a candidate component A formed by multiple local clusters, $n_A = \sum_{a\in A} n_a$ and $\boldsymbol{\mu}_A = n_A^{-1}\sum_{a\in A} n_a\boldsymbol{\mu}_a$, while its element-wise variance is recursively computed as:

$$\mathbf{v}_A = \frac{1}{n_A}\sum_{a\in A} n_a\left[\mathbf{v}_a + (\boldsymbol{\mu}_a - \boldsymbol{\mu}_A)^{\odot 2}\right]. \tag{4}$$

This update decomposes the aggregated dispersion into within-cluster variation and variation among local cluster means. Consequently, candidate components can be constructed and updated entirely from cluster-level statistics, without repeatedly accessing the underlying client samples. We next evaluate cross-client associations through distributional model comparison. Centroid proximity alone provides limited evidence for cross-client association, since local clusters corresponding to the same latent global component may exhibit different dispersions under heterogeneous client distributions. We therefore formulate cross-client association as a statistical model comparison problem. Consider two candidate components A and B, each modeled by a diagonal Gaussian. For component A, the maximized log-likelihood is:

$$\ell(A) = -\frac{n_A}{2}\left[d(1+\log 2\pi) + \sum_{j=1}^{d}\log v_{A,j}\right], \tag{5}$$

where $v_{A,j}$ denotes the empirical variance along dimension j. Let $n = n_A + n_B$, $\omega_A = n_A/n$, and $\omega_B = n_B/n$. With the current memberships of A and B fixed, the separated model has complete-data log-likelihood $\ell_{\text{sep}} = \ell(A) + \ell(B) + n_A\log\omega_A + n_B\log\omega_B$ and $4d+1$ free parameters, whereas the merged model has log-likelihood $\ell(A\cup B)$ and $2d$ free parameters. We quantify their pairwise association by$\Delta\text{BIC}(A,B) = 2\left[\ell_{\text{sep}} - \ell(A\cup B)\right] - (2d+1)\log n$. Smaller values provide stronger statistical evidence for merging the two candidate components.

Assumption 1 (Pairwise Regularity)**.** *The samples represented by candidate components A and B are independently drawn from distributions P_A and P_B with finite second moments and strictly positive marginal variances. Moreover, as $n = n_A + n_B \to \infty$, $n_A/n \to \omega \in (0, 1)$.*

For $P \in \{P_A, P_B\}$, let $G_P^\star \in \mathcal{G}_d$ denote its population-optimal diagonal Gaussian approximation, and let $G_{AB}^\star$ denote the optimal common diagonal Gaussian under the pooled distribution $\omega P_A + (1-\omega)P_B$. We define the corresponding population discrepancy as $\mathcal{D}_\mathcal{G}(P_A, P_B; \omega) = \omega\, \mathbb{E}_{P_A}\left[\log \frac{g_A^\star(\mathbf{x})}{g_{AB}^\star(\mathbf{x})}\right] + (1-\omega)\, \mathbb{E}_{P_B}\left[\log \frac{g_B^\star(\mathbf{x})}{g_{AB}^\star(\mathbf{x})}\right]$, where $g_A^\star$, $g_B^\star$, and $g_{AB}^\star$ denote the densities of $G_A^\star$, $G_B^\star$, and $G_{AB}^\star$, respectively.

Theorem 1 (Asymptotic Characterization of Merge Evidence)**.** *For fixed candidate component memberships A and B, under Assumption 1,*

$$\frac{\Delta\mathrm{BIC}(A, B)}{n} \xrightarrow{p} 2\left[\mathcal{D}_\mathcal{G}(P_A, P_B; \omega) - H(\omega)\right], \tag{6}$$

where $H(\omega) = -\omega \log \omega - (1-\omega)\log(1-\omega)$. Consequently, if $\mathcal{D}_\mathcal{G}(P_A, P_B; \omega) < H(\omega)$, then $\Pr[\Delta\mathrm{BIC}(A, B) < 0] \to 1$; if the inequality is reversed, then $\Pr[\Delta\mathrm{BIC}(A, B) > 0] \to 1$.

Theorem 1 shows that the proposed criterion admits a population-level boundary determined jointly by distributional discrepancy and relative component support. The proof is provided in Appendix A.2.

To avoid mapping multiple local clusters from the same client to one global component, we impose a same-client cannot-link constraint: $\Gamma(A) \cap \Gamma(B) = \varnothing$, where $\Gamma(A)$ denotes the clients represented by candidate component A. This induces the lower bound $K_{\min} = \max_g K_g$. Starting from $M = \sum_{g=1}^{G} K_g$ components, the server iteratively merges the admissible pair with the strongest distributional evidence:

$$(A^\star, B^\star) = \underset{A \neq B:\, \Gamma(A)\cap\Gamma(B)=\varnothing}{\arg\min} \Delta\mathrm{BIC}(A, B). \tag{7}$$

After each merge, the component statistics and affected merge scores are updated recursively. Repeating this procedure yields a nested hierarchy $\mathcal{P}_M \to \cdots \to \mathcal{P}_{K_\mathrm{r}}$, with $K_\mathrm{r} \geq K_{\min}$. Here, $\Delta\mathrm{BIC}$ determines the merge ordering rather than the final model order, which is selected after federated probabilistic optimization.

3.3 Federated Probabilistic Structure Optimization

Candidate structures inferred from local cluster statistics may still be affected by errors introduced during local clustering and hierarchical aggregation. We therefore perform federated probabilistic optimization with sample-level soft assignments to jointly estimate shared global cluster distributions and client-specific cluster compositions. The resulting solutions are then compared to determine the final global cluster number and latent clustering structure. For each candidate $K \in \mathcal{K}_\mathrm{path}$, we model the local data distribution of client g as:

$$p_g(\mathbf{x}) = \sum_{k=1}^{K} \pi_k^{(g)} \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{8}$$

where $p_g(\mathbf{x})$ denotes the local mixture density of client g, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the globally shared mean and covariance matrix of component k, and $\pi_k^{(g)}$ denotes its client-specific mixture weight, satisfying $\pi_k^{(g)} \geq 0$ and $\sum_{k=1}^{K} \pi_k^{(g)} = 1$. Here, $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the corresponding Gaussian component. This formulation shares global information across clients while allowing heterogeneous local cluster compositions. The candidate structure obtained from the first module initializes the mixture model, which is subsequently refined by alternating between client-side expectation and server-side maximization. At communication round t, the server broadcasts the current global component parameters $\{(\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})\}_{k=1}^{K}$. Given these parameters and its local mixture proportions, client g computes the posterior responsibility of global component k for local sample $\mathbf{x}_i$ as:

$$r_{ik}^{(g,t)} = \frac{\pi_k^{(g,t)} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{h=1}^{K} \pi_h^{(g,t)} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_h^{(t)}, \boldsymbol{\Sigma}_h^{(t)}\right)}. \tag{9}$$

These posterior responsibilities provide sample-level soft assignments, allowing the coarse cross-client correspondences obtained during structure discovery to be refined through probabilistic optimization. Instead of transmitting individual responsibilities, client g summarizes its local E-step using three responsibility-weighted sufficient statistics: $N_k^{(g)} = \sum_{i=1}^{n_g} r_{ik}^{(g)}$, $\mathbf{S}_k^{(g)} = \sum_{i=1}^{n_g} r_{ik}^{(g)} \mathbf{x}_i$, and $\mathbf{Q}_k^{(g)} = \sum_{i=1}^{n_g} r_{ik}^{(g)} \mathbf{x}_i \mathbf{x}_i^\top$. Here, $N_k^{(g)}$ represents the effective support of component k, while $\mathbf{S}_k^{(g)}$ and $\mathbf{Q}_k^{(g)}$ summarize its feature statistics and squared-feature statistics. Only these aggregated statistics are communicated to the server, with the raw samples retained locally. The server aggregates the sufficient statistics across clients as $N_k = \sum_{g=1}^{G} N_k^{(g)}$, $\mathbf{S}_k = \sum_{g=1}^{G} \mathbf{S}_k^{(g)}$, and $\mathbf{Q}_k = \sum_{g=1}^{G} \mathbf{Q}_k^{(g)}$. The mean and empirical covariance of global component k are then updated as $\boldsymbol{\mu}_k^{(t+1)} = \mathbf{S}_k / N_k$ and $\widetilde{\boldsymbol{\Sigma}}_k = \mathbf{Q}_k / N_k - \boldsymbol{\mu}_k^{(t+1)} \boldsymbol{\mu}_k^{(t+1)\top}$. Under heterogeneous client distributions, different global components may receive substantially different effective supports. To stabilize covariance estimation for components with limited support, we regularize each component covariance toward a pooled covariance structure:

$$\boldsymbol{\Sigma}_k^{(t+1)} = \alpha_k \widetilde{\boldsymbol{\Sigma}}_k + (1 - \alpha_k) \widetilde{\boldsymbol{\Sigma}}_{\text{pool}}, \qquad \alpha_k = \frac{N_k}{N_k + d}, \tag{10}$$

where $\widetilde{\boldsymbol{\Sigma}}_{\text{pool}}$ denotes the pooled within-component covariance, and α_k adaptively balances component-specific and pooled covariance information based on N_k. Client g then updates its local mixture proportions as $\pi_k^{(g,t+1)} = N_k^{(g)} / n_g$, which preserves heterogeneous local cluster compositions while maintaining globally shared component distributions. The EM steps are repeated until convergence. For candidate K, the observed-data log-likelihood at round t is:

$$\mathcal{L}^{(t)} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \log \left[\sum_{k=1}^{K} \pi_k^{(g,t)} \mathcal{N}\left(\mathbf{x}_i^{(g)}; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)} \right) \right]. \tag{11}$$

Optimization terminates when the relative change in $\mathcal{L}^{(t)}$ falls below a predefined threshold ε, or when the maximum number of federated EM rounds is reached. After federated probabilistic optimization, each $K \in \mathcal{K}_{\text{path}}$ yields an optimized mixture model. We select the final global structure using the Integrated Completed Likelihood (ICL), which balances model fit, complexity, and assignment uncertainty. For candidate K, the Bayesian information criterion is $\text{BIC}_K = -2\mathcal{L}_K + p_K \log N$, where $p_K = Kd + G(K-1) + p_\Sigma(K)$ counts the parameters of the shared means, client-specific mixture proportions, and covariance matrices. We further measure assignment uncertainty using the posterior entropy:

$$H_K = -\sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{k=1}^{K} r_{ik}^{(g)} \log \left(r_{ik}^{(g)} + \epsilon \right). \tag{12}$$

A lower H_K indicates more concentrated posterior responsibilities and therefore higher confidence in cluster assignments, whereas a higher value reflects greater uncertainty among competing components. To incorporate this assignment uncertainty into model selection, we adopt the ICL approximation $\text{ICL}_K = \text{BIC}_K + 2H_K$ and determine the final global cluster number as:

$$\widehat{K} = \arg\min_{K \in \mathcal{K}_{\text{path}}} \text{ICL}_K. \tag{13}$$

Proposition 1 (Consistency of Unknown-K Selection)**.** *Let $\mathcal{K}_0$ be a fixed finite set of admissible model orders such that $\mathcal{K}_{\text{path}} \subseteq \mathcal{K}_0$, and suppose that $\Pr(K^\star \in \mathcal{K}_{\text{path}}) \to 1$. For each $K \in \mathcal{K}_0$, let $\overline{\text{ICL}}_K = \text{ICL}_K / N$ and assume that $\overline{\text{ICL}}_K \xrightarrow{p} \mathcal{J}_K$. If there exists $\delta > 0$ such that:*

$$\mathcal{J}_{K^\star} + \delta < \min_{K \in \mathcal{K}_0 \setminus \{K^\star\}} \mathcal{J}_K, \tag{14}$$

then the selector in Eq. equation 13 satisfies $\Pr(\widehat{K} = K^\star) \to 1$. The proof is provided in Appendix A.3.

The optimized model corresponding to $\widehat{K}$ yields the final global cluster distribution parameters $\widehat{\Theta} = \{(\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k)\}_{k=1}^{\widehat{K}}$. For sample $\mathbf{x}_i$ at client g, the final cluster assignment is obtained by maximum a posteriori inference:

$$\widehat{z}_i^{(g)} = \arg\max_k \left[\log \pi_k^{(g)} + \log \mathcal{N}\left(\mathbf{x}_i; \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k \right) \right]. \tag{15}$$

In summary, FedUGCN first constructs candidate global clustering structures through distribution-aware hierarchical aggregation of local cluster statistics, and then refines each candidate using federated probabilistic optimization with sample-level evidence. By jointly modeling shared global cluster distributions and client-specific cluster compositions, the framework establishes reliable cross-client correspondences while preserving local heterogeneity. The optimized candidates are finally compared using ICL, enabling joint inference of the global cluster number and the corresponding global clustering structure. The overall procedure is summarized in Appendix A.4.

4 EXPERIMENTS

4.1 EXPERIMENT SETUP

Datasets. We conduct experiments on four benchmark datasets: Abalone Nash et al. (1994), FrogA Colonna et al. (2015), Waveform Breiman & Stone (1984), and PenDigits Alpaydin & Alimoglu (1996). FrogA is a clustering variant derived from the Anuran Calls (MFCCs) dataset.

Baselines. We compare our method with seven representative clustering approaches, including GMM McLachlan et al. (2019), k-FED Dennis et al. (2021), FedKmeans Garst & Reinders (2024), FFCM Stallmann & Wilbik (2022), MUFC Pan et al. (2023), AFCL Zhang et al. (2025), and FedGEM Ibrahim et al. (2026). Specifically, GMM serves as a classical probabilistic clustering baseline, while k-FED and FedKmeans represent centroid-based federated clustering methods that construct global structures from local cluster representatives. FFCM extends fuzzy clustering to the federated setting by incorporating soft cluster memberships, whereas MUFC considers federated clustering from the perspective of cluster unlearning. AFCL addresses federated clustering with an unknown number of clusters through adaptive structure discovery, while FedGEM adopts a mixture-modeling framework for inferring global components under unknown model order. Together, these methods cover diverse clustering paradigms for comprehensive comparison.

Implementation Details. For a fair comparison, all methods are evaluated under identical client partitions and random seeds. We consider $G = 10$ clients and adopt a $70\%/30\%$ train–test split. Following the benchmark partition protocol, each client is randomly assigned a subset of the global clusters, resulting in heterogeneous local clustering structures across clients. Samples from the selected clusters are allocated according to their original cluster proportions. Ground-truth labels and the true global cluster number $K^\star$ are used solely for benchmark construction and performance evaluation, and are not involved in model optimization. During optimization, only the partitioned local data and the corresponding local cluster numbers $\{K_g\}_{g=1}^{G}$ are available. We set the number of global communication rounds to 10 and the number of local epochs to 1.

Metrics. Following previous clustering studies Xie et al. (2016); Wang et al. (2025a); Tu et al. (2026), we evaluate clustering performance using clustering accuracy (ACC) and normalized mutual information (NMI), adjusted Rand index (ARI), and F1-score. Higher values indicate better clustering performance. More experimental details are provided in Appendix A.5

4.2 PERFORMANCE COMPARISON

As shown in Table 1, our method demonstrates consistently strong clustering performance across four benchmark datasets under heterogeneous federated settings. In particular, on PenDigits, our method outperforms the second-best method by 15.12%, 13.10%, 21.09%, and 15.42% in terms of ACC, NMI, ARI, and F1, respectively. Similar improvements are observed on Waveform, with gains of 3.14%, 1.57%, 2.57%, and 2.29% across the four metrics. Compared with federated clustering methods such as k-FED, FFCM, MUFC, and FedKmeans, our method achieves more consistent overall performance. These methods construct global structures by aggregating local cluster representatives, which can become unreliable under heterogeneous client distributions and incomplete cross-client observations. In contrast, FedUGCN leverages cluster-level distributional statistics to capture both location and dispersion information, thereby establishing more reliable cross-client cluster associations. Compared with GMM and FedGEM, our framework first identifies plausible candidate global structures and then refines them through federated probabilistic optimization, enabling joint inference of cross-client cluster correspondence and global clustering structure.

Table 1: Clustering performance (%) on four benchmark datasets under a 10-client federated setting. All results are reported as mean ± standard deviation. The best and second-best results are highlighted in **bold** and underlined, respectively.

	Abalone				FrogA			
Method	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1
GMM	32.05±2.12	20.89±0.61	11.68±2.66	30.99±0.54	61.09±8.17	66.61±3.12	47.10±10.79	49.59±3.77
k-FED	45.25±1.30	34.39±2.85	18.41±0.57	43.22±3.88	83.19±7.07	75.98±2.77	75.29±7.43	58.50±3.90
FFCM	43.95±0.58	35.11±3.19	18.82±1.58	46.13±1.03	72.90±4.94	68.68±2.10	67.55±6.11	57.70±9.25
MUFC	44.14±0.96	36.16±2.87	19.20±1.20	47.29±1.26	82.58±1.90	77.73±2.35	76.65±3.50	67.39±3.36
FedKmeans	44.10±1.38	35.25±3.28	18.18±1.51	46.57±1.02	82.32±3.97	76.88±1.00	76.80±6.04	65.90±6.99
AFCL	25.19±0.80	**42.40±3.60**	9.20±1.19	36.52±1.00	58.39±1.71	69.47±3.79	49.18±0.66	73.58±2.96
FedGEM	40.65±4.11	28.37±4.92	15.80±3.34	29.64±5.61	72.29±6.38	63.85±5.06	64.43±8.07	34.12±8.85
FedUGCN	**55.81±1.97**	36.92±2.32	**25.01±0.20**	**53.54±1.28**	**87.43±1.44**	**80.97±2.86**	**77.05±3.58**	**80.33±1.16**

	PenDigits				Waveform			
Method	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1
GMM	61.70±5.07	67.60±0.76	48.91±1.48	59.54±5.58	50.73±0.23	36.11±0.17	25.24±0.14	50.66±0.17
k-FED	68.91±5.29	66.71±2.46	53.34±5.03	69.81±4.98	57.63±2.12	39.40±2.06	27.76±2.66	62.07±3.18
FFCM	52.27±0.07	51.14±2.71	34.89±2.46	46.67±1.43	56.81±0.56	34.31±0.71	26.68±1.22	60.43±2.15
MUFC	70.01±6.30	68.47±4.05	56.17±7.28	72.55±6.62	54.28±1.69	31.30±3.63	23.35±3.25	58.53±2.41
FedKmeans	68.95±4.87	68.92±4.08	55.37±6.53	70.14±5.54	55.44±0.49	38.30±1.22	25.98±1.41	60.44±1.92
AFCL	32.94±1.76	61.34±1.55	27.05±1.85	47.82±2.40	22.76±2.12	35.88±0.63	11.69±0.99	36.32±2.84
FedGEM	53.68±7.45	45.30±3.07	29.51±6.86	50.01±6.83	66.64±1.69	37.43±1.85	31.95±2.28	65.84±1.77
FedUGCN	**85.13±2.67**	**82.02±2.02**	**77.26±3.75**	**87.97±2.44**	**69.78±4.11**	**40.97±5.45**	**34.52±7.78**	**68.13±2.54**

4.3 Unknown-K Estimation

To evaluate the estimation of the unknown global cluster number, we conduct experiments on Abalone and FrogA with 5, 10, 15, and 20 clients. As shown in Figure 2, the proposed method consistently estimates K close to the ground-truth cluster number across different client scales, whereas FedGEM and AFCL exhibit increasingly pronounced overestimation as the number of clients grows. Partitioning the data across more clients leads to increasingly fragmented local clustering structures, making cross-client correspondence and global model-order inference more challenging. The proposed method addresses this issue by recovering cross-client cluster relationships through distribution-aware hierarchical aggregation and selecting feasible global structures via federated probabilistic optimization and ICL. By identifying and merging compatible components across clients, FedUGCN mitigates cluster-number overestimation from fragmented local structures and recovers coherent global relationships, enabling more stable unknown-K inference across varying client scales.

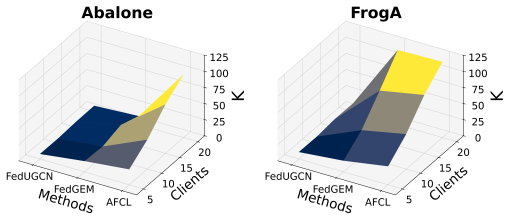

Figure 2: Unknown-K estimation results on Abalone and FrogA under varying numbers of clients. The estimated cluster numbers are reported for different client scales to evaluate the stability of global cluster number estimation.

4.4 Ablation Study

To assess the contributions of the key components in FedUGCN, we compare the full model with "Base", "w/o CSC", and "w/o ACP" on FrogA and PenDigits. As shown in Table 2, FedUGCN achieves the best overall performance on both datasets. Removing CSC reduces ACC by 1.86% and ARI by 1.05% on FrogA, while the corresponding drops on PenDigits increase to 10.37% and 8.92%, respectively, demonstrating the importance of modeling client-specific cluster compositions under heterogeneous distributions. Removing ACP leads to more substantial degradation: ACC decreases by 12.63% and 28.99% on FrogA and PenDigits, while ARI decreases by 16.95% and

Table 2: Ablation study of FedUGCN. "Base" denotes the basic variant without the proposed enhanced components. The variants "w/o CSC" and "w/o ACP" correspond to removing client-specific cluster composition and adaptive covariance pooling. The best result is highlighted in **bold**.

	FrogA				PenDigits			
Method	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1
Base	67.73±0.49	68.80±3.38	52.72±3.10	79.95±3.32	36.33±2.99	62.21±1.52	30.02±1.61	52.24±3.22
w/o CSC	85.57±3.34	78.07±2.28	76.00±3.74	79.21±4.05	74.76±3.61	76.93±1.15	68.34±2.78	82.54±2.22
w/o ACP	74.80±1.86	71.29±2.88	60.10±0.79	77.86±1.74	56.14±2.34	68.04±2.04	50.00±2.32	68.95±0.87
FedUGCN	**87.43±1.44**	**80.97±2.86**	**77.05±3.58**	**80.33±1.16**	**85.13±2.67**	**82.02±2.02**	**77.26±3.75**	**87.97±2.44**

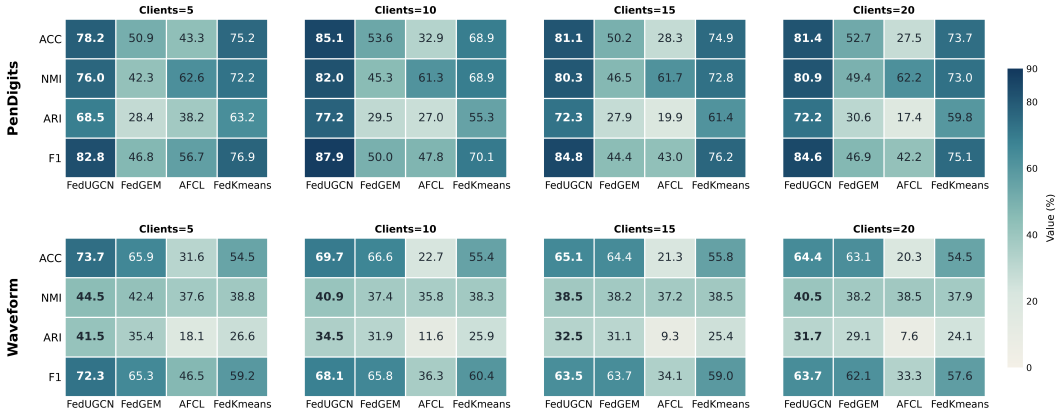

Figure 3: Scalability analysis on PenDigits and Waveform with varying numbers of clients. Clustering performance (%) is evaluated using ACC, NMI, ARI, and F1 under different client scales.

27.26%. These results underscore the role of ACP in stabilizing distribution estimation under uneven cluster support, while CSC complements it by preserving client-specific composition heterogeneity. Their combination therefore enables more reliable and robust federated clustering.

4.5 Scalability Analysis

To evaluate scalability under varying client scales, we conduct experiments on PenDigits and Waveform with different numbers of clients. As shown in Figure 3, the proposed method maintains stable and competitive performance across different client configurations. For example, with 20 clients on PenDigits, it achieves 81.4% ACC, 80.9% NMI, 72.2% ARI, and 84.6% F1, outperforming all competing methods. This robustness stems from distribution-aware global structure discovery, which establishes reliable cross-client cluster associations from local statistics, and federated probabilistic structure optimization, which integrates distributed information to jointly estimate global cluster distributions and client-specific cluster compositions.

5 Conclusion

This paper investigates federated clustering with unknown global cluster numbers under heterogeneous distributed data. To address the challenges of global cluster number estimation and cross-client cluster correspondence, we propose FedUGCN, a probabilistic federated clustering framework for jointly inferring the global cluster number and clustering structure without centralizing raw client data. Specifically, we develop distribution-aware global structure discovery and federated probabilistic structure optimization, which enable reliable cross-client cluster association, candidate global structure construction, and subsequent probabilistic refinement. Extensive experiments demonstrate that FedUGCN achieves competitive clustering performance while effectively inferring the unknown global cluster number across heterogeneous federated settings.

AI USE STATEMENT

In this work, we used generative AI tools to assist with proofreading and improving the clarity and readability of the manuscript. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Ethem Alpaydin and Fevzi Alimoglu. Pen-based recognition of handwritten digits, 1996.

Leo Breiman and Charles J. Stone. Waveform database generator (version 1), 1984.

Xuanyu Chen, Nan Yang, Shuai Wang, and Dong Yuan. Scaling law analysis in federated learning: How to select the optimal model size? In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 20336–20344, 2026a.

Zhichao Chen, Yulong Zhang, Odin Zhang, Fangyikang Wang, Le Yao, Hao Wang, and Zhihuan Song. Blending data and knowledge for process industrial modeling under riemannian preconditioned bayesian framework. *IEEE Transactions on Knowledge and Data Engineering*, pp. 82–95, 2026b.

Yujun Cheng, Weiting Zhang, Zhewei Zhang, Jiawen Kang, Qi Xu, Shengjin Wang, and Dusit Niyato. Snapcfl: A pre-clustering-based clustered federated learning framework for data and system heterogeneities. *IEEE Transactions on Mobile Computing*, 24(6):5214–5228, 2025.

Juan Colonna, Eduardo Nakamura, Marco Cristo, and Marcelo Gordo. Anuran calls (MFCCs), 2015.

Don Kurian Dennis, Tian Li, and Virginia Smith. Heterogeneity for the win: One-shot federated clustering. In *Proceedings of the International Conference on Machine Learning*, pp. 2611–2620, 2021.

Shifei Ding, Chao Li, Lili Guo, and Qiang Yang. Federated clustering: An overview of algorithm evolution and research prospects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):12816–12835, 2026.

Swier Garst and Marcel Reinders. Federated k-means clustering. In *Proceedings of the International Conference on Pattern Recognition*, pp. 107–122, 2024.

Guanxiong He, Zheng Wang, Jie Wang, Liaoyuan Tang, Rong Wang, and Feiping Nie. Towards federated clustering: A client-wise private graph aggregation framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 21619–21627, 2026.

Michael Ibrahim, Nagi Gebraeel, and Weijun Xie. A federated generalized expectation-maximization algorithm for mixture models with an unknown number of components. In *Proceedings of the International Conference on Learning Representations*, 2026.

Longmei Li, ZhenSheng Zang, Nan Zhu, LiYong Zhang, Wei Lu, and Witold Pedrycz. Deep fuzzy c-means clustering in a federated model heterogeneous scenario. *IEEE Transactions on Knowledge and Data Engineering*, 38(8):4958–4971, 2026.

Jiyuan Liu, Xinwang Liu, Siqi Wang, Xinhang Wan, Dongsheng Li, Kai Lu, and Kunlun He. Communication-efficient federated multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):17–32, 2026.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6(1):355–378, 2019.

Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone, 1994.

Andrew A. Neath and Joseph E. Cavanaugh. The bayesian information criterion: background, derivation, and applications. *WIREs Computational Statistics*, 4(2):199–203, 2012.

Chao Pan, Jin Sima, Saurav Prakash, Vishal Rana, and Olgica Milenkovic. Machine unlearning of federated clusters. In *Proceedings of the International Conference on Learning Representations*, 2023.

Morris Stallmann and Anna Wilbik. Towards federated clustering: A federated fuzzy c-means algorithm (ffcm). *arXiv preprint arXiv:2201.07316*, 2022.

David Stenger, Armin Lindicke, Alexander von Rohr, and Sebastian Trimpe. Local entropy search over descent sequences for bayesian optimization. In *Proceedings of the International Conference on Learning Representations*, pp. 16146–16190, 2026.

Andrew A. Stirn and David A. Knowles. The vampprior mixture model. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 1369–1377, 2025.

Jianheng Tang, Yajiang Huang, Kejia Fan, Feijiang Han, Jiaxu Li, Jinfeng Xu, Run He, Anfeng Liu, Houbing Song, Huiping Zhuang, and Yunhuai Liu. Deepafl: Deep analytic federated learning. In *Proceedings of the International Conference on Learning Representations*, pp. 60290–60335, 2026.

Wenxuan Tu, Jieren Cheng, Renxiang Guan, Weixuan Liang, Sihang Zhou, Xin Peng, Zhiping Cai, and Xinwang Liu. Amgc2: Rethinking deep graph clustering with a high attribute-missing ratio. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2026.

Georgios Vardakas, Argyris Kalogeratos, and Aristidis Likas. Uniforce: The unimodality forest method for clustering and estimation of the number of clusters. *Pattern Recognition*, 172:112357, 2026.

Lingren Wang, Wenxuan Tu, Jiaxin Wang, Xiong Wang, Jieren Cheng, and Jingxin Liu. Fedigl: Federated invariant graph learning for non-iid graphs. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 117917–117940, 2025a.

Minglong Wang, Weiwei Li, Yunan Lu, and Xiuyi Jia. Learning label distribution with dirichlet process mixture model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 26398–26406, 2026.

Yizhang Wang, Wei Pang, and Witold Pedrycz. One-shot federated clustering based on stable distance relationships. *IEEE Transactions on Industrial Informatics*, 20(11):13262–13272, 2024.

Yizhang Wang, Wei Pang, Di Wang, and Witold Pedrycz. One-shot secure federated k-means clustering based on density cores. *IEEE Transactions on Neural Networks and Learning Systems*, 36 (8):14131–14143, 2025b.

Yue Wu, Shuaicheng Zhang, Wenchao Yu, Yanchi Liu, Quanquan Gu, Dawei Zhou, Haifeng Chen, and Wei Cheng. Personalized federated learning under mixture of distributions. In *Proceedings of the International Conference on Machine Learning*, pp. 37860–37879, 2023.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the International Conference on Machine Learning*, pp. 478–487, 2016.

Jiaqi Xue, Mayank Kumar, Yuzhang Shang, Shangqian Gao, Rui Ning, Mengxin Zheng, Xiaoqian Jiang, and Qian Lou. Dictpfl: Efficient and private federated learning on encrypted gradients. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 119806–119829, 2025.

Jie Yan, Jing Liu, and Zhong-Yuan Zhang. Omnifc: Rethinking federated clustering via lossless and secure distance reconstruction. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 20819–20847, 2025a.

Shanshan Yan, Zexi Li, Chao Wu, Meng Pang, Yang Lu, Yan Yan, and Hanzi Wang. You are your own best teacher: Achieving centralized-level performance in federated learning under heterogeneous and long-tailed data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2750–2759, 2025b.

Yuan Yao, Lixu Wang, Jiaqi Wu, Jin Song, Simin Chen, Zehua Wang, Zijian Tian, Wei Chen, Huixia Li, and Xiaoxiao Li. Fedre: A representation entanglement framework for model-heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39466–39475, 2026.

Xiaoyang Yi, Li Peng, Yuru Bao, and Jian Zhang. Bayesian evidence-driven prototype evolution for federated domain adaptation. In *Proceedings of the International Conference on Learning Representations*, 2026.

Moxuan Zeng, Wenxuan Tu, Yuanyi Chen, Yiying Wang, Miao Yu, Xiangyan Tang, and Jieren Cheng. Fedpkda: Personalized federated learning with privacy-preserving knowledge dynamic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 28113–28121, 2026.

Li Zhang, Pinhan Fu, Li Lv, Qian Guo, Liang Du, and Xinyan Liang. Evofmvc: Trusted federated multi-view clustering with evolutionary fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 28292–28300, 2026a.

Shaojie Zhang and Ke Chen. Angular constraint embedding via spherepair loss for constrained clustering. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 117065–117119, 2025.

Xu Zhang, Wenpeng Li, Yunfeng Shao, Yonglin Liu, Kaiwen Zhou, and Yinchuan Li. Federated learning via variational bayesian inference: Personalization, sparsity and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):12679–12693, 2026b.

Yue Zhang, Chuanlong Qiu, Xinfa Liao, and Yiqun Zhang. Federated hierarchical clustering with automatic selection of optimal cluster numbers. *Information Sciences*, 733:122957, 2026c.

Yunfan Zhang, Yiqun Zhang, Yang Lu, Mengke Li, Xi Chen, and Yiu-ming Cheung. Asynchronous federated clustering with unknown number of clusters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 22695–22703, 2025.

Tianfei Zhou, Ye Yuan, Binglu Wang, and Ender Konukoglu. Federated feature augmentation and alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11119–11135, 2024.

A Appendix

A.1 Notations

We list the basic notations of this paper in Table 3.

A.2 Proof of Theorem 1

Proof. Under Assumption 1, the empirical means and diagonal variances of A and B converge in probability to their population counterparts. Hence, the fitted diagonal Gaussian models converge to $G_A^\star$ and $G_B^\star$, respectively, while the Gaussian fitted after pooling the two components converges to $G_{AB}^\star$. By the law of large numbers and consistency of these moment estimators, the normalized gain of fitting the two components separately satisfies:

$$\frac{\ell(A)+\ell(B)-\ell(A\cup B)}{n} \xrightarrow{p} \mathcal{D}_{\mathcal{G}}(P_A, P_B; \omega). \tag{16}$$

Meanwhile, since $n_A/n \to \omega$ and $n_B/n \to 1-\omega$,

$$\frac{n_A \log(n_A/n) + n_B \log(n_B/n)}{n} \to -H(\omega). \tag{17}$$

Finally, the BIC complexity term is asymptotically negligible after normalization because $(2d+1)\log n/n \to 0$. Combining these three limits with the definition of $\Delta\mathrm{BIC}(A,B)$ gives Eq. equation 6. If $\mathcal{D}_{\mathcal{G}}(P_A, P_B; \omega) - H(\omega)$ is strictly negative or positive, convergence in probability implies that $\Delta\mathrm{BIC}(A,B)$ eventually has the corresponding sign with probability approaching one. □

Table 3: Summary of the main notations.

Notations	Meaning	Notations	Meaning
$\mathcal{D}_g$	The private unlabeled dataset of client g	$\mathbf{x}_i^{(g)}$	The feature vector of the i-th unlabeled sample of client g
n_g	The number of local samples	d	The feature dimension
N	The total number of samples	$K_g, K^\star, \widehat{K}$	Local/unknown ground-truth global / estimated global number of clusters
$\widehat{\Theta}$	Estimated global cluster distribution parameters	$\mathcal{C}_c^{(g)}$	Local cluster
$\widehat{\boldsymbol{\mu}}_k$	Estimated mean of the k-th global cluster	$\widehat{\boldsymbol{\Sigma}}_k$	Estimated covariance matrix of the k-th global cluster
$n_c^{(g)}$	Local cluster size	$\boldsymbol{\mu}_c^{(g)}$	Local cluster empirical mean
$\mathbf{v}_c^{(g)}$	Local cluster element-wise variance	$\ell(A)$	Maximized log-likelihood of component A
$v_{A,j}$	The empirical variance along dimension j	$\Delta\mathrm{BIC}(A, B)$	Pairwise merge evidence
$\Gamma(A)$	Client set represented by candidate component A	$p_g(\mathbf{x})$	Local mixture distribution of client g
$\pi_k^{(g)}$	Client-specific mixture weight of global cluster k	$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$	Gaussian distribution of global cluster k
t	Communication round	$r_{ik}^{(g,t)}$	Posterior responsibility of component k for sample $\mathbf{x}_i$ at client g
$\pi_k^{(g,t)}$	Client-specific mixture weight at round t	$N_k^{(g)}$	Effective support of component k at client g
$\mathbf{S}_k^{(g)}$	Responsibility-weighted first-order statistic	$\mathbf{Q}_k^{(g)}$	Responsibility-weighted second-order statistic
α_k	The support-dependent coefficient	ICL	The Integrated Completed Likelihood
H_K	the posterior entropy	$\mathcal{L}^{(t)}$	The observed-data log-likelihood at round t
G	Number of clients	$\mathcal{S}_c^{(g)}$	Cluster-level statistics of local cluster c at client g
n_A	Size of candidate component A	$\boldsymbol{\mu}_A$	Empirical mean of candidate component A
$\mathbf{v}_A$	Element-wise variance of candidate component A	$\mathcal{K}_{\mathrm{path}}$	Candidate global cluster number along the hierarchical path
$K_{\min}$	Lower bound of the global cluster number	M	Initial number of local clusters
$\boldsymbol{\mu}_k$	Globally shared mean of component k	$\boldsymbol{\Sigma}_k$	Globally shared covariance matrix of component k
N_k	Aggregated effective support of component k	$\widetilde{\boldsymbol{\Sigma}}_k$	Empirical covariance of component k
$\widetilde{\boldsymbol{\Sigma}}_{\mathrm{pool}}$	Pooled covariance matrix	p_K	Number of free parameters for candidate K
BIC_K	Bayesian information criterion for candidate K	ICL_K	Integrated Completed Likelihood for candidate K
$\mathcal{L}_K$	Optimized log-likelihood for candidate K	$\widehat{z}_i^{(g)}$	Final cluster assignment of sample i at client g

A.3 Proof of Proposition 1

Proof. We first note that the true global cluster number is asymptotically covered by the candidate hierarchy:

$$\Pr(K^\star \in \mathcal{K}_{\mathrm{path}}) \to 1. \tag{18}$$

Therefore, it remains to show that the ICL criterion selects $K^\star$ from the candidate set with probability approaching one. For each $K \in \mathcal{K}_0$, the normalized criterion satisfies $\overline{\mathrm{ICL}}_K \xrightarrow{p} \mathcal{J}_K$. Since $\mathcal{K}_0$ is a fixed finite set of candidate model orders, these pointwise convergences imply:

$$\max_{K \in \mathcal{K}_0} \left| \overline{\mathrm{ICL}}_K - \mathcal{J}_K \right| \xrightarrow{p} 0. \tag{19}$$

Hence, the empirical ICL values uniformly approach their population counterparts over $\mathcal{K}_0$. Let $\epsilon \in (0, \delta/2)$ and define the event:

$$\mathcal{E}_N = \left\{ \max_{K \in \mathcal{K}_0} \left| \overline{\mathrm{ICL}}_K - \mathcal{J}_K \right| < \epsilon \right\}. \tag{20}$$

By Eq. equation 19, $\Pr(\mathcal{E}_N) \to 1$. On this event, for any $K \in \mathcal{K}_0 \setminus \{K^\star\}$,

$$\begin{aligned} \overline{\mathrm{ICL}}_K - \overline{\mathrm{ICL}}_{K^\star} &\geq \mathcal{J}_K - \mathcal{J}_{K^\star} - 2\epsilon \\ &\geq \delta - 2\epsilon > 0, \end{aligned} \tag{21}$$

where the second inequality follows from the separation condition in Eq. equation 14. Thus, the positive population-level gap between $K^\star$ and every competing model order in $\mathcal{K}_0$ is preserved by the empirical ICL criterion with probability approaching one.

Since $\mathcal{K}_{\mathrm{path}} \subseteq \mathcal{K}_0$, whenever $K^\star \in \mathcal{K}_{\mathrm{path}}$ and $\mathcal{E}_N$ occurs, $K^\star$ is also the unique minimizer over the candidate hierarchy:

$$\arg\min_{K \in \mathcal{K}_{\mathrm{path}}} \overline{\mathrm{ICL}}_K = K^\star. \tag{22}$$

Since division by the common positive factor N does not change the minimizer, the same conclusion holds for ICL_K itself.

Algorithm 1 FedUGCN: Probabilistic Federated Clustering with Unknown Global Cluster Number

Require: Local datasets $\{\mathcal{D}_g\}_{g=1}^{G}$; local cluster numbers $\{K_g\}_{g=1}^{G}$; maximum communication rounds T

Ensure: Estimated global cluster number $\widehat{K}$; global parameters $\widehat{\Theta}$; client-specific mixture proportions $\{\widehat{\boldsymbol{\pi}}^{(g)}\}_{g=1}^{G}$

for each client $g = 1, \ldots, G$ **in parallel do**
  Perform local clustering with K_g clusters
  Compute cluster-level statistics $\{\mathcal{S}_c^{(g)}\}_{c=1}^{K_g}$
  Upload the cluster-level statistics to the server
end for
Initialize $M = \sum_{g=1}^{G} K_g$ candidate components
Set $K_{\min} = \max_g K_g$ and $\mathcal{K}_{\text{path}} \leftarrow \{M\}$
while the number of components is larger than $K_{\min}$ **do**
  Construct admissible pairs under the same-client cannot-link constraint
  if no admissible pair exists **then**
    break
  end if
  Evaluate pairwise merge evidence for all admissible pairs
  Select the pair to merge according to Eq. equation 7
  Add the resulting model order to $\mathcal{K}_{\text{path}}$
end while
for each $K \in \mathcal{K}_{\text{path}}$ **do**
  Initialize the probabilistic model in Eq. equation 8 from the candidate structure
  for $t = 1, \ldots, T$ **do**
    Server broadcasts the current global parameters
    for each client $g = 1, \ldots, G$ **in parallel do**
      Compute posterior responsibilities using Eq. equation 9
      Compute local sufficient statistics $\{N_k^{(g)}, \mathbf{S}_k^{(g)}, \mathbf{Q}_k^{(g)}\}_{k=1}^{K}$
      Upload the sufficient statistics to the server
    end for
    Aggregate client sufficient statistics
    Update the globally shared component means
    Update component covariances using Eq. equation 10
    Update client-specific mixture proportions
    if the relative change in Eq. equation 11 is below the convergence threshold **then**
      break
    end if
  end for
  Compute posterior entropy using Eq. equation 12
  Compute ICL_K
end for
Select $\widehat{K}$ using Eq. equation 13
return $\widehat{K}$, $\widehat{\Theta}$, $\{\widehat{\boldsymbol{\pi}}^{(g)}\}_{g=1}^{G}$

Combining Eqs. equation 18 and equation 19, we obtain:

$$\begin{aligned} \Pr(\widehat{K} = K^\star) &\geq \Pr(K^\star \in \mathcal{K}_{\text{path}}, \mathcal{E}_N) \\ &\geq 1 - \Pr(K^\star \notin \mathcal{K}_{\text{path}}) - \Pr(\mathcal{E}_N^c) \to 1. \end{aligned} \tag{23}$$

Therefore, the ICL-based selector is consistent for the unknown global cluster number under the stated candidate-coverage and separation conditions. □

A.4 Algorithmic Procedure of FedUGCN

Algorithm 1 summarizes the overall procedure of FedUGCN.

A.5 EXPERIMENTS

A.5.1 DATASETS

We evaluate our method on four benchmark datasets. The dataset statistics are summarized in Table 4. **1) Abalone** Nash et al. (1994) is a real-world dataset containing physical measurements of abalone specimens; following our experimental protocol, we use seven numerical features and eight benchmark clusters, with substantial overlap among cluster distributions. **2) FrogA** Colonna et al. (2015) is derived from the Anuran Calls (MFCCs) dataset and contains 7,195 acoustic samples represented by 21 Mel-frequency cepstral coefficient features across 10 clusters, exhibiting complex and partially overlapping distributions. **3) PenDigits** Alpaydin & Alimoglu (1996) is a handwritten digit benchmark with 16 numerical features extracted from pen trajectories and 10 digit classes, providing a relatively well-separated yet heterogeneous clustering scenario. **4) Waveform** Breiman & Stone (1984) contains 5,000 samples with 21 continuous features generated from three underlying waveform classes, whose overlapping distributions make it a widely used benchmark for evaluating clustering robustness.

Table 4: Dataset statistics.

Dataset	Samples	Dim.	#Clusters
Abalone	4,177	7	8
FrogA	7,195	21	10
Waveform	5,000	21	3
PenDigits	10,992	16	10

A.5.2 BASELINES

We compare FedUGCN with representative centralized and federated clustering approaches, covering both known- and unknown-K settings.

- **GMM McLachlan et al. (2019).** Gaussian Mixture Model is a centralized probabilistic clustering baseline that models the data distribution as a mixture of Gaussian components. The global cluster number is provided in advance.
- **k-FED Dennis et al. (2021).** k-FED is a one-shot federated clustering method in which clients perform local K-means clustering and communicate local cluster centers to the server for global aggregation. It assumes prior knowledge of the global cluster number.
- **FFCM Stallmann & Wilbik (2022).** Federated Fuzzy C-Means extends fuzzy clustering to the federated setting by coordinating local fuzzy cluster information and aggregating local cluster representatives to construct global clusters. It requires the global cluster number to be specified beforehand.
- **MUFC Pan et al. (2023).** MUFC is a one-shot federated clustering framework based on local cluster representatives, designed to support clustering under heterogeneous and imbalanced client distributions while enabling federated cluster unlearning.
- **FedKmeans Garst & Reinders (2024).** FedKmeans extends K-means to the federated setting through iterative client–server coordination for updating global cluster centers. Similar to conventional K-means, it assumes that the global cluster number is known.
- **AFCL Zhang et al. (2025).** AFCL addresses federated clustering without a prespecified global cluster number by initializing an excessive number of candidate centers and adaptively determining the resulting cluster structure through competitive and balancing mechanisms.
- **FedGEM Ibrahim et al. (2026).** FedGEM is a probabilistic federated clustering method for an unknown global cluster number. Clients perform local EM optimization and construct uncertainty sets for local components, while the server identifies potential cross-client component overlaps to infer the global cluster structure.

A.5.3 SCALABILITY ANALYSIS

Figure 4 evaluates the robustness of FedUGCN under varying client scales. On Abalone, FedUGCN maintains strong clustering performance as the number of clients increases from 5 to 20, with ACC improving from 46.2% to 53.2% and NMI from 27.2% to 43.4%. On FrogA, FedUGCN exhibits stronger scalability, achieving ACC values of 87.4% and 88.1% with 10 and 20 clients, respectively,

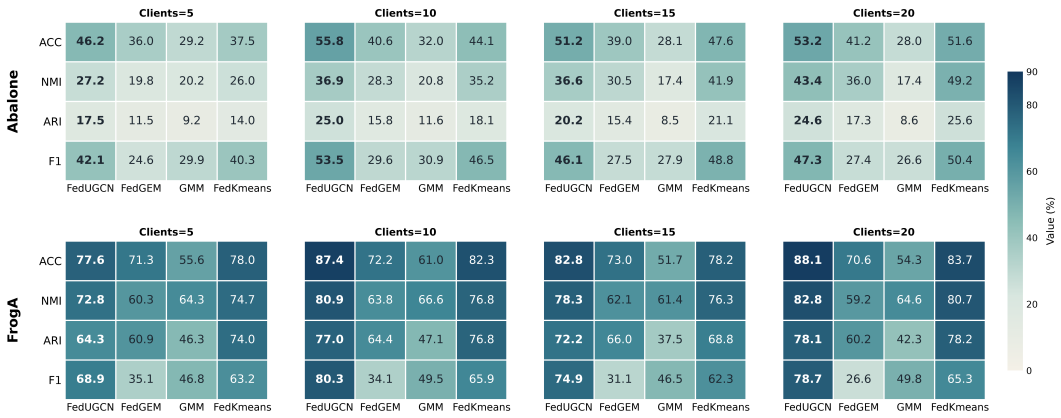

Figure 4: Scalability analysis on Abalone and FrogA with varying numbers of clients. Clustering performance (%) is evaluated using ACC, NMI, ARI, and F1 under different client scales.

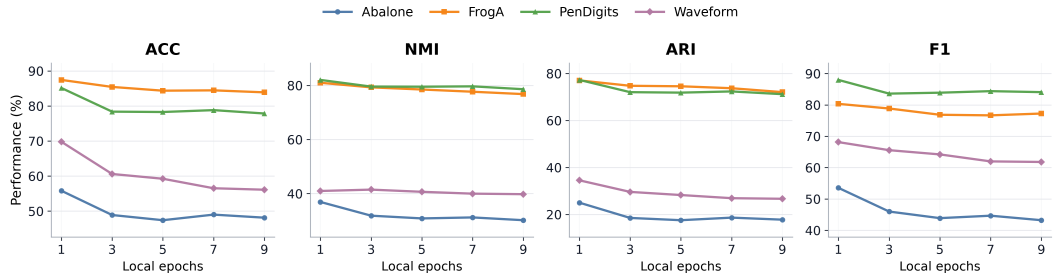

Figure 5: Effect of varying local training epochs on FedUGCN across four benchmark datasets in terms of ACC, NMI, ARI, and F1.

while maintaining competitive performance in terms of NMI, ARI, and F1. The stable performance across both datasets demonstrates the robustness of FedUGCN to increasing client scales, enabling reliable global clustering structure inference in more distributed federated settings.

A.5.4 Effect of Local Training Epochs

Figure 5 investigates the effect of local training epochs on FedUGCN. Overall, fewer local updates lead to better clustering performance, with most datasets and evaluation metrics achieving their best or near-best results at a single local epoch. As the number of local epochs increases, performance degradation becomes more pronounced on Abalone and Waveform, whereas FrogA and PenDigits remain relatively stable across a wider range of local training settings. This suggests that the sensitivity to local optimization intensity varies with the underlying data distribution and cluster structure. These results indicate that excessive local updates may cause local solutions to become overly adapted to client-specific distributions, thereby amplifying cross-client discrepancies and weakening the consistency of subsequent global structure modeling. In contrast, a small number of local updates is sufficient to capture informative local structure while limiting client drift. Consequently, FedUGCN can maintain stable global clustering structure inference with relatively low local computational overhead, which is desirable for resource-constrained federated settings.