FEDN-TAGC: FEDERATED NODE-LEVEL TEXT-ATTRIBUTED GRAPH CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Node-level federated graph clustering studies how several holders of unlabeled subgraphs can train clustering models without sharing local graph data. Existing methods usually rely on graph topology or vector attributes, but many real graphs also attach texts to nodes. In text-attributed graphs, a node's text and its local neighborhood may point to different clusters, which can corrupt local cluster evidence and cross-client semantic alignment. We propose **FedN-TAGC**, a Federated Node-Level Text-Attributed Graph Clustering method for this setting. On each client, FedN-TAGC learns text-view, structure-view, and fused node representations, and compares text-view and structure-view assignments under the same fused cluster centers to separate stable samples from structure-text conflict cases. Stable samples are summarized by an LLM into local cluster descriptions, while conflict cases are assigned after the server forms global topics. The server groups uploaded summaries into global topics, aligns client clusters, and sends semantic consensus back to guide later local clustering. Experiments on benchmark text-attributed graphs show that this conflict-handling design improves federated node clustering.

1 INTRODUCTION

Federated Graph Learning (FGL) lets multiple data holders learn from distributed graphs without centralizing raw nodes, edges, or attributes (Fu et al., 2022; He et al., 2021). This setting appears in academic networks, social platforms, recommendation systems, and biomedical graphs, where records are often owned by different organizations or devices. While much of FGL focuses on supervised prediction, graph clustering is needed when node labels are expensive, incomplete, or unavailable (Wang et al., 2019; Bo et al., 2020). Many such graphs also attach texts to nodes, such as paper abstracts, user profiles, item descriptions, or clinical notes. We study *Federated Node-Level Text-Attributed Graph Clustering* (FedN-TAGC), where clients cluster nodes from distributed text-attributed subgraphs while keeping local graph data private.

Graph clustering has first been studied in centralized settings, where deep clustering, graph auto-encoding, and attributed graph clustering methods learn latent groups from features and graph structure without labels (Kipf & Welling, 2020; Sun et al., 2020; Wang et al., 2019; Bo et al., 2020). Federated graph clustering adds another constraint: each client sees only a subgraph, so the local graph may miss neighbors, edges, or attributes that would be available in a centralized graph. Recent work has therefore studied missing neighbors, client-specific subgraph distributions, domain shift, cross-subgraph link repair, and incomplete attributes in FGL (Baek et al., 2023; Wan et al., 2024; Liu et al., 2025; 2026). FedSage+ and FedNCN show that graph partition can damage local structural evidence, while IFedGC shows that missing attributes also change clustering behavior. These methods show that local clustering is shaped by the information boundary and data availability of each client. Still, they mainly work with ordinary attributed graphs, where node attributes are treated as vectors and the semantics behind the attributes are not examined directly.

TAGs introduce another source of information: node texts carry topic semantics, while neighborhoods reflect citation, interaction, or co-occurrence patterns. Graph neural networks such as GCN and GAT provide standard tools for structural aggregation (Kipf & Welling, 2016; Velickovic et al., 2018), while recent TAG benchmarks and TAG foundation models show that language models and graph models extract different signals from text and structure (Yan et al., 2023; Chen et al., 2024b; He et al., 2025; Zhou et al., 2025; Zhang et al., 2024). This distinction matters more in a federated

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.

setting, because each client observes only part of the graph but still needs to produce cluster evidence that can be compared with other clients. Recent TAG-FGL methods use large language models for missing-neighbor generation, link inference, efficient condensation, and multimodal defense (Yan et al., 2025; Chen et al., 2026; Shi et al., 2026). They show how LLMs can provide semantic signals for TAG-FGL, but their training and evaluation still rely mainly on supervised node classification. Moving from supervised TAG-FGL to unsupervised federated TAG clustering is therefore not a direct extension, especially when reliable labels are scarce, costly to annotate, or unavailable across clients. More details on centralized graph clustering, subgraph FGL, and TAG-FGL are provided in Appendix A.

In this setting, two obstacles arise. First, without node labels, supervised TAG-FGL solutions lose the label signals or label anchors used for training and cross-client alignment (Yan et al., 2025; Chen et al., 2026; Shi et al., 2026). The server cannot use labels to match clusters across clients, so local cluster descriptions become the main objects that it can compare. Second, as shown in Figure 1, text and structure may give conflicting evidence for the same node. This is related to the broader observation that graph neighborhoods are not always semantically homogeneous, and GNNs can fail when connected nodes carry different labels or features (Zhu et al., 2020; Gong et al., 2023; Pan & Kang, 2023). For TAGs, the same issue can appear as a text-structure mismatch: a node can be semantically close to one latent topic while its local neighborhood supports another (Yan et al., 2023; Chen et al., 2024b; Zhou et al., 2025). If such nodes are summarized too early, the local descriptions sent to the server may mix different topics, be treated as global topics, and return to clients as misleading consensus. The problem is amplified by privacy constraints, because the server can inspect only uploaded summaries rather than raw texts or local edges. These obstacles make supervised TAG-FGL methods and ordinary federated graph clustering methods insufficient for this problem.

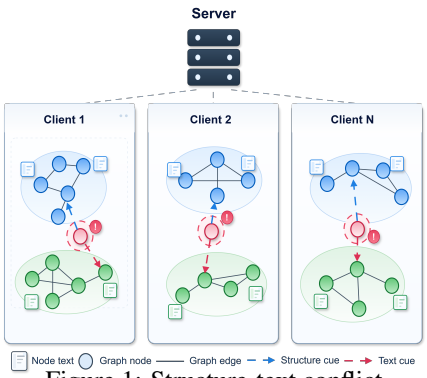

Figure 1: Structure-text conflict.

We propose **FedN-TAGC**, a Federated Node-Level Text-Attributed Graph Clustering method with local conflict screening and global semantic consensus. On each client, FedN-TAGC learns a graph-propagated representation for local clustering and a text-only diagnostic representation, then detects uncertain nodes by comparing their assignments under the same cluster centers. Stable samples are used as semantic anchors and summarized by an LLM into local cluster descriptions, while high-conflict nodes remain pending until later semantic resolution. On the server, the LLM groups uploaded summaries into global topics, aligns client clusters, and updates semantic consensus. The returned consensus gives globally aligned topic descriptions for later local clustering and pending-node assignment. This design links the two parts of the problem: local summaries are built from more stable evidence, and pending nodes are judged after local clusters have been placed in a shared semantic space. FedN-TAGC can therefore reduce structure-text noise while keeping the task unsupervised and decentralized. In summary, our contributions are threefold:

- **New Task Focus.** We formulate FedN-TAGC as an unsupervised federated node-level clustering setting for text-attributed graphs, and identify structure-text conflict as a source of unreliable local summaries and cross-client misalignment.
- **Conflict Handling.** We design a semantic pipeline that detects text-structure assignment disagreement, builds LLM-based summaries from stable evidence, and uses local-global semantic consensus to resolve pending nodes.
- **Evaluation.** We evaluate FedN-TAGC on benchmark text-attributed graphs and compare it with representative federated and graph clustering baselines, showing that explicit conflict handling improves federated node clustering.

2 METHODOLOGY

We present **FedN-TAGC** for unsupervised federated node-level clustering on text-attributed graphs (Yan et al., 2023; Zhou et al., 2025). The method is built around a simple observation: before clients

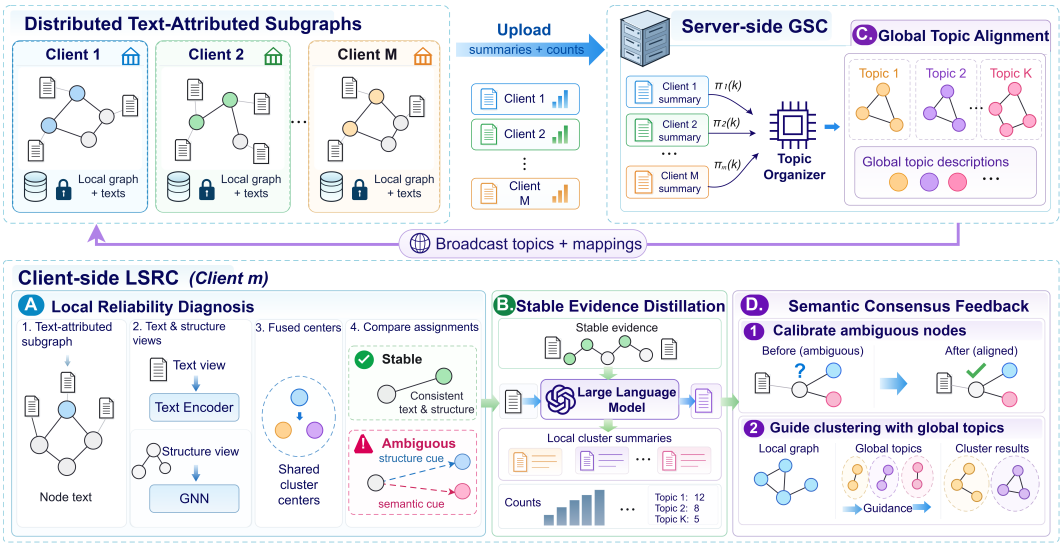

Figure 2: Overall framework of FedN-TAGC. Each client performs local semantic reliability calibration by comparing text-view and structure-view assignments under shared clustering centers, distills stable evidence into local semantic summaries, and sends only summaries and counts to the server. The server builds a global semantic topic space, aligns local clusters through topic mappings, and broadcasts topic consensus for local clustering and semantic calibration.

describe their clusters to the server, they should first decide which local nodes provide stable evidence and which nodes carry divergent text and structure signals. If ambiguous nodes are summarized too early, their texts may pull a local cluster description toward one topic while their graph neighborhoods support another, making server-side topic alignment unreliable.

As shown in Figure 2, FedN-TAGC contains two coupled modules. The *Local Semantic Reliability Calibration module (LSRC)* runs on each client. It learns a graph-propagated representation for local clustering, contrasts it with a text-only diagnostic representation, stratifies nodes by cross-view reliability, and distills local cluster semantics from stable evidence. Ambiguous nodes are retained as semantic calibration candidates. The *Global Semantic Consensus module (GSC)* runs on the server. It organizes uploaded local summaries into global topics, aligns local cluster IDs to these topics, and returns topic-level consensus for local clustering and semantic calibration.

This organization makes semantic reasoning explicit in the federated clustering pipeline. LSRC decides which local evidence is sufficiently coherent to describe a cluster, whereas GSC converts these locally reliable descriptions into a shared topic coordinate system. The returned consensus then acts as a semantic reference for both cluster alignment and the interpretation of cross-view ambiguous nodes.

2.1 Problem Formulation

We consider M clients. Client m holds a local text-attributed subgraph

$$\mathcal{G}_m = (\mathcal{V}_m, \mathcal{E}_m, \mathbf{X}_m, \mathcal{T}_m), \tag{1}$$

where $\mathcal{V}_m, \mathcal{E}_m$ are local nodes and edges, $\mathbf{X}_m \in \mathbb{R}^{n_m \times d}$ contains text-derived node features, and $\mathcal{T}_m = \{t_i\}_{i \in \mathcal{V}_m}$ contains raw node texts. No node labels are available during training. The goal is to assign all local nodes to K latent clusters through communication rounds without sending raw texts, node identities, local edges, dense node embeddings, disagreement scores, or labels to the server.

Let $\mathbf{A}_m$ denote the normalized local adjacency matrix and $\mathbf{I}_m$ the identity adjacency. A local encoder f_{θ_m} maps text-derived features and graph structure to node representations, following standard graph neural aggregation mechanisms (Kipf & Welling, 2016; Velickovic et al., 2018). FedN-TAGC uses the graph-propagated output as the clustering representation and compares it with a text-only diagnostic

pass:

$$\mathbf{Z}_m^s = f_{\theta_m}(\mathbf{X}_m, \mathbf{A}_m), \qquad \mathbf{Z}_m^t = f_{\theta_m}(\mathbf{X}_m, \mathbf{I}_m), \qquad \mathbf{Z}_m \equiv \mathbf{Z}_m^s. \tag{2}$$

Here $\mathbf{Z}_m^s$ includes graph-neighborhood information and is also the representation used for clustering, denoted by $\mathbf{Z}_m$ for the rest of the paper. The text-view representation $\mathbf{Z}_m^t$ uses the same encoder with identity adjacency, so it emphasizes text semantics without neighborhood propagation. This notation reflects the practical implementation: FedN-TAGC does not introduce an additional fusion encoder; it diagnoses structure-text disagreement by contrasting the graph-propagated clustering view with the text-only pass under the same cluster centers.

2.2 Local Semantic Reliability Calibration Module

Local clustering backbone. Each client first trains its encoder with a compact clustering backbone loss (Xie et al., 2016; Bo et al., 2020). The decoder reconstructs the locally smoothed features $\mathbf{A}_m\mathbf{X}_m$ as $\bar{\mathbf{X}}_m$. Given cluster centers $\mathbf{C}_m = \{\mathbf{c}_{m,k}\}_{k=1}^K$, the clustering representation produces a Student-t soft assignment $\mathbf{Q}_m$ (Xie et al., 2016); as in deep embedded clustering, $\mathbf{Q}_m$ is aligned with a sharpened target distribution $\mathbf{P}_m$. We write the backbone loss as

$$\mathcal{L}_{\text{base}}^m = \left\|\mathbf{A}_m\mathbf{X}_m - \bar{\mathbf{X}}_m\right\|_F^2 + \mathrm{KL}(\mathbf{P}_m \| \mathbf{Q}_m). \tag{3}$$

After local optimization, client m runs K-means (MacQueen et al., 1967) on $\mathbf{Z}_m$ and obtains temporary local assignments $y_{m,i}$ and clustering centers $\mathbf{C}_m$. These centers provide the common reference used to compare text-view and structure-view evidence.

Using shared clustering centers as the reference avoids judging the two diagnostic views in separate coordinate systems. A node is considered reliable only when its textual semantics and propagated neighborhood signal support the same cluster. Thus, the following stratification is not a heuristic filtering step; it is a local semantic consistency test grounded in the clustering geometry.

Cross-view reliability stratification. For node i on client m, FedN-TAGC computes the nearest clustering center under the text view and the structure view:

$$a_{m,i}^t = \arg\min_k \left\|\mathbf{z}_{m,i}^t - \mathbf{c}_{m,k}\right\|_2^2, \qquad a_{m,i}^s = \arg\min_k \left\|\mathbf{z}_{m,i}^s - \mathbf{c}_{m,k}\right\|_2^2. \tag{4}$$

Agreement, $a_{m,i}^t = a_{m,i}^s$, indicates that the node's text and local neighborhood point to the same cluster. Disagreement indicates a possible cross-view semantic mismatch, a phenomenon also observed in text-attributed graph learning and non-homophilous neighborhoods (Yan et al., 2023; Chen et al., 2024b; Zhu et al., 2020). FedN-TAGC further scores the strength of disagreement:

$$\begin{aligned} r_{m,i} = \; & \max\{d_{m,i}^t(a_{m,i}^s) - d_{m,i}^t(a_{m,i}^t), 0\} + \max\{d_{m,i}^s(a_{m,i}^t) - d_{m,i}^s(a_{m,i}^s), 0\} \\ & + \frac{1}{2}\left(\Delta_{m,i}^t + \Delta_{m,i}^s\right), \end{aligned} \tag{5}$$

where $d_{m,i}^t(k) = \|\mathbf{z}_{m,i}^t - \mathbf{c}_{m,k}\|_2^2$, $d_{m,i}^s(k) = \|\mathbf{z}_{m,i}^s - \mathbf{c}_{m,k}\|_2^2$, and $\Delta_{m,i}^t, \Delta_{m,i}^s$ are nearest-second-nearest margins in the two views. The first two terms measure how much one view rejects the other view's preferred cluster, and the margin terms keep low-confidence disagreements from dominating the calibration set.

Within each local cluster, only high-tail disagreements enter the semantic calibration set:

$$b_{m,i} = \mathbb{I}\left[a_{m,i}^t \neq a_{m,i}^s \;\wedge\; r_{m,i} > \tau_{m,y_{m,i}}\right], \tag{6}$$

where $\tau_{m,y_{m,i}}$ is the adaptive high-tail threshold computed inside cluster $y_{m,i}$. Nodes outside this set are treated as reliable evidence for cluster semantics, while nodes inside it are reserved for semantic calibration under the global topic space. This reliability stratification is performed before any semantic evidence is uploaded.

Semantic evidence distillation and calibration. Let $\mathcal{S}_{m,k} = \{i \mid y_{m,i} = k,\ b_{m,i} = 0\}$ be the stable evidence set for local cluster k. The client selects central and diverse texts from $\mathcal{S}_{m,k}$ and asks the LLM to produce a compact local summary $s_{m,k}$, in line with recent LLM-assisted federated TAG learning (Yan et al., 2025; Chen et al., 2026). In our implementation, this summarizer is GPT-5 accessed through a Responses-API-compatible endpoint. Calibration candidates are not used in this summary. The server therefore receives only

$$u_{m,k} = (m, k, s_{m,k}, n_{m,k}), \tag{7}$$

where $n_{m,k} = |\mathcal{S}_{m,k}|$ is the amount of stable evidence supporting the summary. Raw texts and node-level representations remain local.

After global topics are available, FedN-TAGC performs semantic arbitration for the calibration set. For each node whose local views disagree, the resolver compares its text with the global topic descriptions and assigns a calibrated topic membership; the resolver output is schema-validated before use:

$$\mathcal{C}_m[i] = g_i, \qquad i \in \mathcal{B}_m, \tag{8}$$

where $\mathcal{B}_m = \{i \mid b_{m,i} = 1\}$ and g_i is the selected global topic. In this way, stable nodes define local semantic evidence, and ambiguous nodes are interpreted in the shared topic space instead of by local cluster indices alone.

Local view-consistency loss. LSRC also adds a consistency term on stable nodes only:

$$\mathcal{L}_{\mathrm{st}}^{m} = \frac{1}{|\mathcal{S}_m| + \epsilon} \sum_{i \in \mathcal{S}_m} \left[\mathrm{KL}(\mathbf{q}_{m,i} \| \mathbf{q}_{m,i}^{t}) + \mathrm{KL}(\mathbf{q}_{m,i} \| \mathbf{q}_{m,i}^{s}) \right], \tag{9}$$

where $\mathcal{S}_m = \cup_k \mathcal{S}_{m,k}$, and $\mathbf{q}_{m,i}$, $\mathbf{q}_{m,i}^{t}$, and $\mathbf{q}_{m,i}^{s}$ are cluster distributions computed from the clustering, text-view, and structure-view representations with the same centers. Because the loss is applied only to stable nodes, it strengthens consistent evidence without forcing detected ambiguous nodes into an unreliable local assignment.

2.3 Global Semantic Consensus Module

The server receives the stable local summaries $\mathcal{U} = \{u_{m,k} \mid m = 1, \ldots, M,\ k = 1, \ldots, K\}$ and builds a global semantic topic space. In the first communication round, the LLM groups all summaries into K global topics,

$$\mathcal{H} = \{h_1, \ldots, h_K\}, \tag{10}$$

and assigns each local summary to a topic. Topic discovery, topic updates, and summary-to-topic alignment are accepted only when they match the required JSON schemas. Across communication rounds, the topic identities are maintained as a global semantic coordinate system: new local summaries are aligned to existing topics, and the topic descriptions are updated from their aligned evidence. The assignment is many-to-one, allowing a client to miss a global topic or contain multiple local clusters related to the same topic.

The global semantic consensus is obtained at the summary level rather than by averaging node embeddings. Specifically, the server compares local summaries by their described concepts, merges summaries that express the same latent topic, and rewrites each topic description from all aligned stable evidence. The count $n_{m,k}$ serves as evidence strength, so summaries supported by more stable nodes have a stronger influence on the final topic description without exposing the underlying raw nodes.

Let $\pi_m(k)$ denote the global topic aligned to local cluster k on client m. The server broadcasts the topic descriptions and the local-to-global mapping π_m. Clients then use this consensus in two ways. First, as described in Eq. 8, calibration candidates are assigned by comparing their texts with the global topics. Second, the topic descriptions provide a weak semantic anchor for local clustering. Let $\phi(\cdot)$ be the text embedding function used to encode a topic description, and let $\bar{\mathbf{x}}_{m,k}$ be the mean normalized text feature of stable nodes in local cluster k. FedN-TAGC defines

$$\mathcal{L}_{\mathrm{gsc}}^{m} = \frac{1}{K} \sum_{k=1}^{K} \left[1 - \cos\left(\bar{\mathbf{x}}_{m,k}, \phi(h_{\pi_m(k)}) \right) \right]. \tag{11}$$

This term does not send raw node texts to the server. It only encourages the client's stable cluster evidence to remain compatible with the returned global topic semantics. If no global topics have been formed yet, this term is disabled.

The server may additionally train a FedNCN-style proxy model (Liu et al., 2025) from transformed proxy features and a proxy graph derived from selected local representatives. This feedback supplies model and cluster-layer initialization, while the semantic consensus remains the mechanism that aligns clusters and calibrates ambiguous nodes. Labels are never used for representation learning, reliability diagnosis, summary generation, topic alignment, consensus update, or semantic calibration.

2.4 Overall Objective and Training

Algorithm 1 summarizes how LSRC and GSC are alternated across communication rounds before defining the optimization objective.

The client-side objective in round r is

$$\mathcal{L}_m = \mathcal{L}^m_{\text{base}} + \lambda_{\text{st}}\mathcal{L}^m_{\text{st}} + \lambda_{\text{gsc}}\gamma_r\mathcal{L}^m_{\text{gsc}}, \quad (12)$$

where $\gamma_r = \mathbb{I}(|\mathcal{H}| > 0)$ disables global semantic guidance before the first server update. The total objective contains three loss blocks: the local clustering backbone $\mathcal{L}^m_{\text{base}}$, the stable view-consistency loss $\mathcal{L}^m_{\text{st}}$, and the global semantic consensus loss $\mathcal{L}^m_{\text{gsc}}$. FedN-TAGC keeps only two tunable loss weights: λ_{st} for stable local view consistency and λ_{gsc} for global semantic consensus.

Algorithm 1: Procedure of FedN-TAGC

Require: $\{\mathcal{G}_m\}_{m=1}^M$, cluster number K, rounds R.
Require: Weights $\lambda_{\text{st}}, \lambda_{\text{gsc}}$.
Ensure: Clustering result RS.
1: Initialize global topics $\mathcal{H} \leftarrow \emptyset$.
2: **for** $r = 0, \ldots, R-1$ **do**
3: **Local Semantic Reliability Calibration module (LSRC).**
4: **for** client m **do**
5: Optimize local encoder by Eq. 12.
6: Run KMeans on $\mathbf{Z}_m$ to obtain $\mathbf{C}_m, y_m$.
7: Select calibration set $\mathcal{B}_m$ by Eq. 6.
8: Upload stable summaries $\{u_{m,k}\}_{k=1}^K$.
9: **end for**
10: **Global Semantic Consensus module (GSC).**
11: Align uploaded summaries to global topics.
12: Update topic descriptions $\mathcal{H}$.
13: Broadcast $\mathcal{H}$ and local-to-global mappings.
14: **for** client m **do**
15: Calibrate $\mathcal{B}_m$ by Eq. 8.
16: **end for**
17: **end for**
18: Return RS from aligned client assignments.

2.5 Communication and Raw-Data Non-Leakage

Each client uploads at most K natural-language summaries per round, one for each non-empty local cluster with stable evidence, together with the corresponding evidence counts. If proxy-model feedback is enabled, it also uploads transformed proxy features and a proxy graph summary, following the same raw-data non-leakage rule. The server never receives raw node texts, raw text features, node identities, local edges, dense node embeddings, disagreement scores, or labels. Representative texts may be sent to the trusted LLM endpoint for semantic processing, but they are not part of the federated server payload or saved server audit artifacts. The additional communication of the semantic path is therefore $\mathcal{O}(MKL_s)$ per round in the worst case, where L_s is the average summary length. This cost is independent of the number of local nodes except through the preset number of clusters, making the semantic alignment stage lightweight compared with node-level representation exchange.

3 Experiments

3.1 Experimental Setup

Benchmark Datasets. We evaluate FedN-TAGC on five public text-attributed graph benchmarks (Yan et al., 2023; Zhou et al., 2025): Cora, PubMed, WikiCS, CiteSeer, and Wisconsin. Nodes correspond to documents or web pages, edges describe citation or hyperlink relations, and node attributes are derived from the associated texts. Cora, PubMed, and CiteSeer use 384-dimensional SBERT text features from the MARK text-attributed graph benchmark source, while WikiCS and Wisconsin provide additional web-page benchmarks with different graph scales and homophily patterns. Dataset statistics are reported in Appendix D.

Federated Partition Settings. Each benchmark graph is partitioned into $M \in \{5, 10, 20\}$ clients. The three client scales simulate increasingly fragmented node ownership and stronger cross-client distribution shift. We use topology-aware partitioning to construct induced local subgraphs, so each client can access only its own nodes, local edges, node texts, and text features. Cross-client edges are removed from local message passing and are not sent to the server. Node labels are never used for partitioning, local training, semantic summarization, global topic discovery, cluster alignment, or server aggregation; they are used only after training for evaluation. More details of the federated protocol are given in Appendix E.

Baseline Methods. The main comparison includes two unsupervised federated graph clustering baselines, FedNCN (Liu et al., 2025) and IFedGC (Liu et al., 2026), and four federated graph learning methods adapted to the unsupervised protocol, FedPUB-U, FedTAD-U (Zhu et al., 2024), FedGTA-U

Table 1: Main comparison under different datasets and client settings. -U denotes supervised methods adapted to the unsupervised protocol. **Bold** and underlined values denote the best and second-best performance, respectively.

Dataset	Models	Client 5				Client 10				Client 20			
		ACC	NMI	ARI	F1	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1
Cora	FedNCN	47.67±1.00	30.38±4.41	19.75±1.88	40.09±3.89	50.95±0.23	32.23±4.28	28.88±7.53	35.97±10.55	34.84±12.66	15.00±16.98	10.28±11.12	22.78±19.27
	IFedGC	47.06±0.76	30.41±5.64	22.30±3.57	32.09±9.75	38.13±2.74	30.34±0.71	19.16±0.23	35.06±3.05	42.52±3.73	15.63±15.65	9.01±11.21	25.10±17.59
	FedPUB-U	12.52±0.46	8.41±0.36	7.36±0.53	8.11±0.15	6.96±0.28	11.97±0.16	4.06±0.57	1.45±0.27	28.91±0.74	8.10±0.01	3.87±0.81	16.80±0.01
	FedTAD-U	13.34±0.71	0.42±0.31	0.88±0.35	6.93±1.10	5.98±0.23	9.45±0.01	4.33±0.51	0.78±0.09	31.19±0.73	4.78±0.21	2.01±0.75	11.49±0.16
	FedGTA-U	12.74±0.46	4.73±0.17	5.39±0.01	8.06±0.41	24.09±0.32	6.42±0.88	7.30±0.01	16.50±0.41	17.69±0.13	3.38±0.71	7.20±0.01	11.07±0.80
	FedIIH-U	25.76±1.39	0.74±1.49	0.46±0.67	14.51±1.89	18.37±0.01	3.99±0.67	0.31±0.60	10.09±0.66	15.84±0.01	5.81±0.69	2.36±0.28	12.60±0.89
	Ours	**62.72±1.02**	**44.00±0.05**	**38.34±1.58**	**57.64±3.49**	**62.92±2.35**	**43.98±0.47**	**40.24±1.30**	**59.04±1.17**	**64.49±0.39**	**42.02±0.73**	**36.71±1.55**	**62.46±0.50**
PubMed	FedNCN	56.07±1.00	16.98±0.01	15.09±2.21	49.69±1.03	52.18±0.78	11.37±0.21	9.49±1.87	40.56±0.33	52.68±1.00	10.21±0.10	10.67±0.32	40.38±0.00
	IFedGC	55.17±2.50	16.87±0.74	15.35±2.82	43.72±2.89	48.33±2.20	7.28±0.74	4.41±1.32	40.73±8.27	47.91±4.70	4.07±0.66	3.50±1.38	34.50±0.84
	FedPUB-U	41.54±0.75	3.66±0.90	3.05±0.33	21.74±0.01	31.27±0.14	4.58±0.51	4.85±0.16	26.15±0.88	41.69±2.15	3.19±0.32	3.68±0.23	35.44±1.29
	FedTAD-U	39.75±0.41	7.68±0.23	1.78±0.29	25.29±0.81	27.97±0.99	2.94±0.09	3.72±0.54	33.06±0.87	31.57±0.30	2.86±0.28	2.79±0.34	24.77±0.25
	FedGTA-U	28.51±0.01	3.50±0.14	4.03±0.80	36.77±0.43	31.81±0.33	7.86±0.01	7.89±0.43	25.06±0.30	33.68±0.01	4.22±0.98	2.43±0.74	24.02±0.23
	FedIIH-U	45.61±0.50	1.43±1.73	0.62±1.16	33.81±2.46	42.59±0.95	4.35±0.78	3.14±0.18	20.95±0.01	26.61±1.78	2.12±0.65	7.79±0.40	20.23±0.59
	Ours	**62.44±1.26**	**18.16±0.34**	**19.99±0.60**	**61.83±1.31**	**61.67±2.59**	**17.97±2.08**	**19.80±5.86**	**60.66±2.69**	**58.12±3.18**	**13.61±0.51**	**14.71±3.39**	**57.38±2.32**
WikiCS	FedNCN	36.81±0.00	23.58±0.00	22.07±0.00	21.91±0.00	46.05±1.41	34.27±5.81	23.89±1.48	36.44±3.22	41.15±2.53	28.39±2.63	17.70±4.89	33.33±0.10
	IFedGC	47.22±0.44	31.22±1.53	24.48±1.39	35.82±0.48	47.32±8.23	17.52±14.22	14.02±8.78	20.74±15.66	35.14±4.42	25.29±1.67	13.84±3.15	26.38±0.02
	FedPUB-U	11.14±0.01	0.44±0.35	2.86±0.37	5.23±0.19	8.65±0.80	0.70±1.02	2.82±0.23	4.25±0.90	8.71±0.12	1.94±0.43	2.44±0.15	4.32±0.10
	FedTAD-U	8.62±0.13	0.46±0.51	2.60±0.27	4.72±0.17	8.09±0.32	0.15±0.12	4.74±0.20	4.64±0.16	7.94±0.21	0.97±0.30	5.04±0.34	3.12±1.18
	FedGTA-U	7.96±0.60	0.75±0.31	1.74±0.34	7.39±0.69	8.03±0.93	0.61±0.16	5.22±0.86	4.77±0.19	7.68±0.51	3.97±0.01	0.03±0.33	3.06±0.74
	FedIIH-U	23.25±0.50	1.90±2.40	0.67±1.64	9.86±0.87	13.44±0.01	2.11±0.61	0.38±0.58	3.69±0.79	11.38±0.26	1.11±0.49	0.55±0.18	6.22±0.01
	Ours	**48.99±0.50**	**34.85±1.65**	**26.72±0.72**	**37.34±0.39**	**59.34±1.88**	**41.41±0.96**	**41.57±0.22**	**48.55±2.56**	**55.59±1.93**	**41.70±0.59**	**37.65±2.17**	**44.30±2.88**
CiteSeer	FedNCN	44.92±0.00	24.06±0.00	20.84±0.00	31.77±0.00	40.77±0.00	19.06±0.00	12.87±0.00	27.81±0.00	38.91±0.21	14.30±0.15	8.15±1.57	24.80±1.14
	IFedGC	46.38±1.08	25.59±0.68	21.29±0.66	31.95±1.43	45.92±0.04	23.37±0.01	18.06±0.02	28.35±0.04	36.98±1.03	14.37±1.41	11.51±0.96	32.08±1.25
	FedPUB-U	23.15±1.22	0.44±1.04	7.33±0.37	7.52±0.29	23.83±0.98	9.89±0.07	4.98±0.35	21.30±0.77	13.13±0.01	9.20±0.23	6.95±0.16	9.91±0.40
	FedTAD-U	18.44±0.42	5.95±0.27	2.51±0.52	12.56±0.27	26.13±0.70	6.37±0.23	2.41±0.18	17.50±1.23	10.29±0.33	8.47±0.18	3.02±0.95	7.36±0.62
	FedGTA-U	13.24±0.17	4.88±0.85	3.56±0.11	21.95±0.39	13.07±0.01	5.50±0.80	4.02±0.88	6.09±0.35	26.07±0.79	9.15±0.58	2.98±0.18	21.35±0.77
	FedIIH-U	24.30±0.35	4.45±1.64	0.86±0.21	18.25±0.93	17.44±0.32	2.80±0.37	0.13±0.01	14.46±0.98	19.30±0.58	3.84±0.22	0.09±0.95	13.62±0.05
	Ours	**59.01±3.86**	**31.06±1.83**	**29.29±3.66**	**52.24±3.20**	**53.12±1.53**	**26.63±1.58**	**25.28±1.71**	**47.69±0.39**	**51.05±1.80**	**25.32±1.20**	**23.73±1.74**	**46.22±0.96**
Wisconsin	FedNCN	50.75±8.80	30.09±11.62	22.14±13.07	38.83±6.36	57.55±2.93	33.49±3.39	26.82±1.77	39.11±2.99	33.96±26.15	18.48±4.63	16.98±14.14	22.59±11.22
	IFedGC	44.34±10.95	21.83±7.16	24.71±25.17	27.53±10.47	41.13±22.42	26.12±4.12	24.12±15.30	29.89±14.76	34.15±26.95	22.66±10.63	24.48±22.29	26.30±18.93
	FedPUB-U	33.41±1.60	4.44±0.84	3.63±0.53	14.08±0.09	17.33±0.18	4.26±0.32	3.06±0.87	24.65±1.79	17.04±0.21	5.14±0.44	7.71±0.87	28.07±1.66
	FedTAD-U	13.82±0.99	19.49±1.20	4.32±1.07	22.98±1.19	21.52±0.08	4.07±0.24	6.43±0.86	29.01±1.23	19.76±0.01	6.04±0.26	1.75±0.89	24.76±1.36
	FedGTA-U	31.34±1.53	1.31±0.77	7.14±0.08	9.35±0.42	32.91±1.16	4.71±0.96	2.41±0.31	16.24±0.50	29.94±0.89	7.05±1.33	3.12±0.10	16.37±0.93
	FedIIH-U	15.78±0.82	15.97±1.54	2.65±0.96	26.95±0.74	30.50±1.36	8.59±1.08	2.83±0.34	13.73±0.81	32.41±0.91	4.70±1.08	5.95±0.01	13.50±0.84
	Ours	**71.51±0.80**	**46.26±2.79**	**54.90±2.90**	**51.88±3.90**	**66.79±1.60**	**49.42±0.68**	**49.39±5.20**	**50.80±3.93**	**66.79±0.53**	**43.63±1.24**	**48.89±1.03**	**51.68±0.69**

(Li et al., 2024), and FedIIH-U. FedNCN is the closest topology-driven federated node-clustering baseline, while IFedGC represents recent incomplete federated graph clustering. For FedPUB, FedTAD, FedGTA, and FedIIH, label-based objectives or classifier heads are removed or disabled, and the learned representations are clustered for post-hoc evaluation; the suffix '-U' denotes these unsupervised adaptations. The aggregation ablation further keeps the same local clustering backbone and replaces only the collaboration strategy with Local and three generic aggregation baselines: FedAvg, FedProx, and FedPer (McMahan et al., 2017; Li et al., 2020; Arivazhagan et al., 2019). This isolates whether generic federated aggregation can explain the gains of FedN-TAGC. Baseline details are provided in Appendix F.

Implementation Details. All methods are run under the same dataset partitions, client scales, and evaluation protocol whenever applicable. FedN-TAGC uses a FedNCN-style graph autoencoder backbone (Liu et al., 2025), with a graph-propagated latent clustering representation and a text-only diagnostic pass for reliability estimation. Unless otherwise specified, clients optimize local models with Adam and perform K-means (MacQueen et al., 1967) on the learned representations; the server receives only compact semantic summaries or proxy information required by the corresponding protocol, rather than raw node texts, raw features, local edges, node identifiers, or labels. The convergence analysis tracks communication rounds explicitly, while the main comparison uses the final configuration described in Appendix G. The semantic operations use GPT-5 through a Responses-API-compatible endpoint; prompt limits and validation rules are reported in the appendix.

Evaluation Metrics. We report clustering accuracy (ACC), normalized mutual information (NMI) (Strehl & Ghosh, 2002), adjusted Rand index (ARI) (Hubert & Arabie, 1985), and macro-F1. ACC and macro-F1 are computed after optimal cluster-to-label matching, and all labels are hidden during training. Higher values indicate better clustering quality for all metrics. Results are reported as mean and standard deviation over repeated runs and multiplied by 100. Metric definitions are given in Appendix H.

3.2 Comparison Experiments

Table 1 compares FedN-TAGC with all baselines under different datasets and client scales. FedN-TAGC achieves the best results on all reported metrics, showing stable clustering quality when

node ownership becomes increasingly fragmented. FedNCN and IFedGC are the strongest competitors, but their lower NMI and ARI indicate weaker global cluster consistency after local graphs are partitioned across clients. This gap is especially meaningful because NMI and ARI are more sensitive to whether local cluster assignments are globally coherent. The supervised federated graph methods adapted to the unsupervised protocol are less competitive, suggesting that classification-oriented objectives and label-dependent training signals cannot directly resolve cross-client semantic mismatch. The gains on both citation and web-page graphs further show that FedN-TAGC is robust to different feature distributions, graph homophily patterns, and client scales.

3.3 Federated Aggregation Ablation

We further replace only the collaboration rule with Local, FedAvg, FedProx, and FedPer to test whether generic aggregation explains the gains. The local encoder, clustering loss, client partition, and evaluation protocol are kept unchanged, so the comparison isolates the effect of server-side collaboration. Local removes communication and serves as the client-only lower bound, while FedAvg, FedProx, and FedPer represent averaging, drift-regularized aggregation, and personalized aggregation, respectively. This design separates ordinary model aggregation from the proposed cluster-level semantic collaboration. Table 2 reports Client 5 results, and Client 10/20 results are in Appendix I. FedAvg is often strong on ACC, while FedPer and FedProx occasionally improve other metrics, but all generic aggregation variants remain below FedN-TAGC because they aggregate parameters rather than align cluster semantics across clients. In particular, parameter aggregation cannot explicitly identify reliable local semantic evidence or resolve the mismatch between local cluster IDs and global topics. This confirms that the main gain comes from semantic collaboration instead of simply changing the federated optimizer.

Table 2: Aggregation strategy ablation on Client 5.

Dataset	Strategy	ACC	NMI	ARI	F1
Cora	Local	49.30±3.32	26.48±2.00	17.90±3.91	41.19±4.78
	FedAvg	49.90±0.41	0.14±0.15	0.94±0.43	9.74±0.35
	FedProx	17.22±0.15	0.87±0.13	0.21±0.88	10.88±0.10
	FedPer	20.20±0.01	0.93±0.71	0.86±0.87	11.05±0.28
	Ours	**62.72±1.02**	**44.00±0.05**	**38.34±1.58**	**57.64±3.49**
PubMed	Local	48.11±3.56	1.55±2.24	2.35±2.70	45.98±3.71
	FedAvg	48.46±0.99	3.64±0.24	6.27±0.60	30.02±0.75
	FedProx	0.79±0.57	0.76±0.30	0.30±0.33	0.38±0.01
	FedPer	36.38±1.50	3.60±0.69	2.52±0.17	28.05±0.62
	Ours	**62.44±1.26**	**18.16±0.34**	**19.99±0.60**	**61.83±1.31**
WikiCS	Local	35.43±2.70	17.06±4.43	6.12±3.02	20.60±2.19
	FedAvg	38.57±0.22	6.33±0.45	6.25±0.29	13.68±0.25
	FedProx	7.73±0.65	0.19±0.30	0.29±0.27	4.44±0.60
	FedPer	4.71±0.01	3.47±0.33	4.00±0.14	9.67±0.10
	Ours	**48.99±0.50**	**34.85±1.65**	**26.72±0.72**	**37.34±0.39**
CiteSeer	Local	43.70±4.35	13.16±4.08	10.36±5.45	35.60±4.65
	FedAvg	49.44±0.08	6.33±0.29	3.30±0.75	9.20±0.17
	FedProx	21.30±0.52	1.28±0.50	0.32±0.33	14.74±0.72
	FedPer	28.15±0.87	8.84±0.33	3.33±0.19	23.60±0.27
	Ours	**59.01±3.86**	**31.06±1.83**	**29.29±3.66**	**52.24±3.20**
Wisconsin	Local	56.97±2.93	28.47±5.59	35.33±5.14	35.84±4.72
	FedAvg	57.53±2.51	6.96±1.70	0.77±0.01	16.14±2.95
	FedProx	27.80±4.05	12.41±3.30	0.39±2.13	20.50±5.04
	FedPer	26.11±4.08	12.32±3.55	0.82±2.80	20.05±4.56
	Ours	**71.51±0.80**	**46.26±2.79**	**54.90±2.90**	**51.88±3.90**

3.4 Module Ablation

Figure 3 evaluates LSRC and GSC by removing each module or both. LSRC selects reliable local semantic evidence from text-view and structure-view assignments, while GSC converts uploaded local evidence into a shared global semantic reference. Removing both modules leads to the largest drop, since the model loses both reliable semantic filtering and global alignment. Removing LSRC mainly weakens local evidence quality, while removing GSC weakens cross-client consistency because clients lack a shared topic-level reference. The single-removal variants may differ across datasets because local view conflict and client heterogeneity vary across graphs. The full model performs best, indicating that LSRC and GSC solve complementary parts of the federated clustering problem; more results are shown in Appendix J.

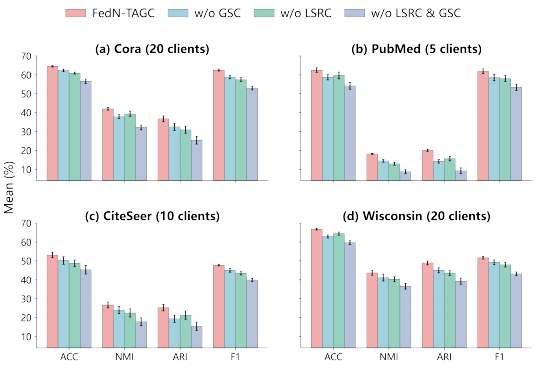

Figure 3: Module ablation.

3.5 Convergence Analysis

Figure 4 tracks FedN-TAGC across communication rounds, with the remaining curves in Appendix J. Each round updates local representations, uploads semantic summaries, constructs global consensus, and returns the consensus to clients. The performance improves quickly in early rounds, suggesting that the first several rounds already provide useful cross-client semantic correction. Small fluctuations appear before convergence because local reliable evidence and global topics are updated together, and the server consensus can slightly change as better summaries are collected. After the semantic topics become stable, the curves remain close to the final performance. This indicates that FedN-TAGC can obtain stable clusters without excessive communication, and that semantic consensus acts as a training signal rather than a post-processing step.

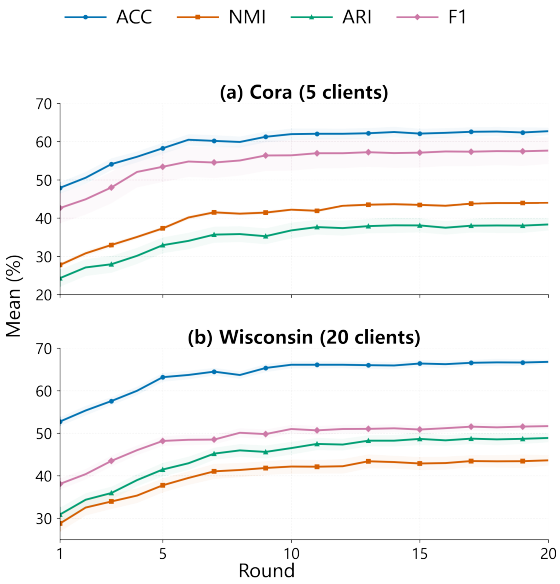

Figure 4: Convergence curves.

3.6 Hyperparameter Analysis

Figure 5 studies the two loss weights λ_{st} and λ_{gsc} over $\{0.01, 0.1, 1, 10, 100\}$. The former controls stable local view consistency, and the latter controls global semantic consensus. These two terms affect different stages of FedN-TAGC: λ_{st} suppresses unreliable local assignments, while λ_{gsc} adjusts how strongly clients follow the returned global topics. The balanced setting usually gives the best or near-best ACC, while nearby settings remain competitive, suggesting that FedN-TAGC is not overly sensitive to exact hyperparameter values. Very small weights may provide insufficient semantic regularization, whereas overly large weights may over-constrain local clustering before representations become stable; additional results are provided in Appendix J.

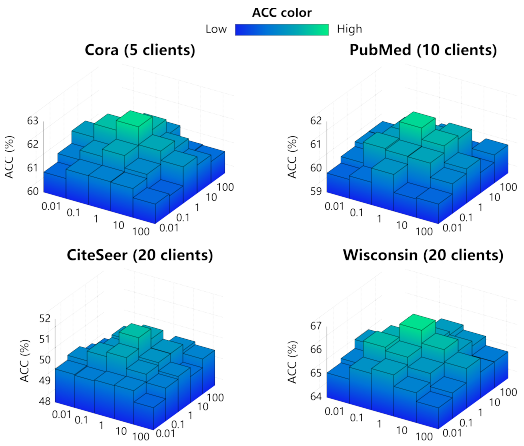

Figure 5: Sensitivity to λ_{st} and λ_{gsc} on ACC.

4 Conclusion

We introduced FedN-TAGC, an unsupervised federated node-level clustering framework for text-attributed graphs. The central idea is that local clusters should not be summarized before checking whether their textual and structural evidence are mutually reliable. FedN-TAGC therefore separates stable nodes from structure-text conflict cases on each client, distills local cluster semantics from stable evidence, and uses server-side global semantic consensus to align client clusters and calibrate ambiguous nodes without exposing raw texts, local edges, node identities, dense embeddings, or labels to the federated server. Experiments on five benchmark text-attributed graphs under multiple client partitions show that this conflict-aware semantic collaboration consistently improves clustering quality over federated graph clustering baselines and generic aggregation strategies. Future work will study stronger privacy mechanisms for natural-language summaries, more stable LLM-based topic alignment, and extensions to dynamic graphs and evolving client topic coverage.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, improving the readability of the manuscript, LaTeX formatting, and debugging author-written code. They were not used to develop the proposed methodology, design the experiments, formulate or prove mathematical claims, process the datasets, or interpret the experimental results. All AI-assisted content was carefully reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Javad Aliakbari, Johan Östman, and Alexandre Graell i Amat. Decoupled subgraph federated learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/78c30547255e433e5ef8ffd88f114765-Abstract-Conference.html.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya K Singh, and Sunpreet S Arora. 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 *International Conference on Machine Learning*, pp. 1396–1415, 2023.

Deyu Bo, Xiao Wang, Chuan Shi, Meiqi Zhu, Emiao Lu, and Peng Cui. Structural Deep Clustering Network. In *Proceedings of the Web Conference 2020*, pp. 1400–1410, 2020.

Mulin Chen, Bocheng Wang, and Xuelong Li. Deep Contrastive Graph Learning with Clustering-Oriented Guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11364–11372, 2024a.

Zekai Chen, Haodong Lu, Xunkai Li, Henan Sun, Jia Li, Hongchao Qin, Rong-Hua Li, and Guoren Wang. Dance: Dynamic, available, neighbor-gated condensation for federated text-attributed graphs. *arXiv preprint arXiv:2601.16519*, 2026. URL https://arxiv.org/abs/2601.16519.

Zhikai Chen, Haitao Mao, Jingzhe Liu, Yu Song, Bingheng Li, Wei Jin, Bahare Fatemi, Anton Tsitsulin, Bryan Perozzi, Hui Liu, et al. Text-space Graph Foundation Models: Comprehensive Benchmarks and New Insights. *Advances in Neural Information Processing Systems*, 37:7464–7492, 2024b.

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

Lele Fu, Bowen Deng, Sheng Huang, Tianchi Liao, Shirui Pan, and Chuan Chen. Less is more: Federated graph learning with alleviating topology heterogeneity from a causal perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17727–17745. PMLR, 2025a. URL https://proceedings.mlr.press/v267/fu25a.html.

Xingbo Fu, Bingrui Zhang, Yushun Dong, Chen Chen, and Jundong Li. Federated graph machine learning: A survey of concepts, techniques, and applications. *ACM SIGKDD Explorations Newsletter*, 24(2):32–47, 2022. doi: 10.1145/3575637.3575644. URL https://doi.org/10.1145/3575637.3575644.

Yiwei Fu, Yuxing Zhang, Chunchun Chen, Jianwen Ma, Quan Yuan, Rong-Cheng Tu, Xinli Huang, Wei Ye, Xiao Luo, and Minghua Deng. MARK: Multi-agent collaboration with ranking guidance for text-attributed graph clustering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6057–6072, Vienna, Austria, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.314. URL https://aclanthology.org/2025.findings-acl.314/.

Lei Gong, Sihang Zhou, Wenxuan Tu, and Xinwang Liu. Attributed Graph Clustering with Dual Redundancy Reduction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3015–3021, 2022.

Zheng Gong, Guifeng Wang, Ying Sun, Qi Liu, Yuting Ning, Hui Xiong, and Jingyu Peng. Beyond Homophily: Robust Graph Anomaly Detection via Neural Sparsification. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2104–2113, 2023.

Chaoyang He, Keshav Balasubramanian, Emir Ceyani, Carl Yang, Han Xie, Lichao Sun, Lifang He, Liangwei Yang, S Yu Philip, Yu Rong, et al. Fedgraphnn: A federated Learning Benchmark System for Graph Neural Networks. In *ICLR 2021 Workshop on Distributed and Private Machine Learning (DPML)*, 2021.

Yufei He, Yuan Sui, Xiaoxin He, and Bryan Hooi. Unigraph: Learning a Unified Cross-domain Foundation Model for Text-attributed Graphs. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 448–459, 2025.

Lawrence Hubert and Phipps Arabie. Comparing Partitions. *Journal of Classification*, 2:193–218, 1985.

Sungwon Kim, Yoonho Lee, Yunhak Oh, Namkyeong Lee, Sukwon Yun, Junseok Lee, Sein Kim, Carl Yang, and Chanyoung Park. Subgraph federated learning for local generalization. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/83b17fb3369b1effa97ca5409526b02e-Abstract-Conference.html.

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Thomas N. Kipf and Max Welling. Graph Auto-Encoders. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. In *Proceedings of the Machine Learning and Systems*, pp. 429–450, 2020.

Xunkai Li, Zhengyu Wu, Wentao Zhang, Yinlin Zhu, Rong-Hua Li, and Guoren Wang. Fedgta: Topology-aware Averaging for Federated Graph Learning. *ArXiv Preprint ArXiv:2401.11755*, 2024.

Jingxin Liu, Renda Han, Wenxuan Tu, Haotian Wang, Junlong Wu, and Jieren Cheng. Federated node-level clustering network with cross-subgraph link mending. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38540–38556. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liu25v.html.

Jingxin Liu, Haotian Wang, Renda Han, Yaoming Wang, Jianxiong Huang, Zhen Zhang, Xiqi Zhao, Zhanyu Ma, and Jieren Cheng. Causally-aware attribute completion for incomplete federated graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19316–19324, 2026. doi: 10.1609/aaai.v40i18.39547. URL https://ojs.aaai.org/index.php/AAAI/article/view/39547.

James MacQueen et al. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14): 281–297, 1967.

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.

Erlin Pan and Zhao Kang. Beyond Homophily: Reconstructing Structure for Graph-agnostic Clustering. In *Proceedings of the International Conference on Machine Learning*, volume 202, pp. 26868–26877, 2023.

Tong Shi, Yilun Jin, Yu Wang, Haowei Wang, Kai Chen, Hanjun Dai, Lu Yu, Bolun Li, Xiangnan He, and Wenqi Zhang. Towards robust text-attributed federated graph learning via unveiling multimodal attack and defense. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20467–20475, 2026. doi: 10.1609/aaai.v40i19.39679. URL https://ojs.aaai.org/index.php/AAAI/article/view/39679.

Alexander Strehl and Joydeep Ghosh. Cluster Ensembles–a Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3(Dec):583–617, 2002.

Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. In *International Conference on Learning Representations (ICLR)*, 2020.

Zihan Tan, Guancheng Wan, Wenke Huang, He Li, Guibin Zhang, Carl Yang, and Mang Ye. Fedspa: Generalizable federated graph learning under homophily heterogeneity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15464–15475, 2025.

Puja Trivedi, Nurendra Choudhary, Edward W. Huang, Vassilis N. Ioannidis, Karthik Subbian, and Danai Koutra. Large language model guided graph clustering. In *The Third Learning on Graphs Conference*, 2024. URL https://openreview.net/forum?id=CLyhlb5DG5.

Petar Velickovic, Guillem Cucurull, Alberto Casanova, Javier Romero, Pietro Lio, and Yoshua Bengio. Graph Attention Networks. In *Proceedings of the International Conference on Learning Representations*, 2018.

Guancheng Wan, Wenke Huang, and Mang Ye. Federated graph learning under domain shift with generalizable prototypes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 15429–15437, 2024.

Chun Wang, Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, and Chengqi Zhang. Attributed Graph Clustering: a Deep Attentional Embedding Approach. In *International Joint Conference on Artificial Intelligence 2019*, pp. 3670–3676, 2019.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised Deep Embedding for Clustering Analysis. In *International Conference on Machine Learning*, pp. 478–487, 2016.

Bo Yan, Zhongjian Zhang, Huabin Sun, Mengmei Zhang, Yang Cao, and Chuan Shi. Data-centric federated graph learning with large language models. *arXiv preprint arXiv:2503.19455*, 2025. URL https://arxiv.org/abs/2503.19455.

Hao Yan, Chaozhuo Li, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, Weiwei Deng, Qi Zhang, Lichao Sun, Xing Xie, and Senzhang Wang. A comprehensive study on text-attributed graphs: Benchmarking and rethinking. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17238–17264. Curran Associates, Inc., 2023. doi: 10.52202/075280-0755. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/37d00f567a18b478065f1a91b95622a0-Paper-Datasets_and_Benchmarks.pdf.

Ke Zhang, Carl Yang, Xiaoxiao Li, Lichao Sun, and Siu Ming Yiu. Subgraph federated learning with missing neighbor generation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6671–6682, 2021. URL https://arxiv.org/abs/2106.13430.

Mengmei Zhang, Mingwei Sun, Peng Wang, Shen Fan, Yanhu Mo, Xiaoxiao Xu, Hong Liu, Cheng Yang, and Chuan Shi. Graphtranslator: Aligning Graph Model to Large Language Model for Open-ended Tasks. In *Proceedings of the ACM Web Conference*, pp. 1003–1014, 2024.

Chuang Zhou, Zhu Wang, Shengyuan Chen, Jiahe Du, Qiyuan Zheng, Zhaozhuo Xu, and Xiao Huang. Taming language models for text-attributed graph learning with decoupled aggregation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3463–3474, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.173. URL https://aclanthology.org/2025.acl-long.173/.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond Homophily in Graph Neural Networks: Current Limitations and Effective Designs. *Advances in Neural Information Processing Systems*, 33:7793–7804, 2020.

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. *ArXiv Preprint ArXiv:2404.14061*, 2024.

A Related Work

A.1 Node-Level Text-Attributed Graph Clustering

Node-level graph clustering aims to partition nodes into latent clusters without using node labels. Early deep clustering methods jointly learn latent representations and cluster assignments through self-training objectives (Xie et al., 2016). Attributed graph clustering further incorporates graph structure into the clustering process. DAEGC jointly models node attributes and neighborhood structure to learn representations for clustering (Wang et al., 2019), while SDCN integrates structural information into deep clustering by combining an autoencoder with a graph convolutional network (Bo et al., 2020). S3GC further studies scalable self-supervised graph clustering without label supervision (Devvrit et al., 2022). AGC-DRR reduces redundant information in both the input and latent spaces for attributed graph clustering (Gong et al., 2022), and DCGL introduces clustering-oriented guidance into contrastive graph learning (Chen et al., 2024a). These methods jointly model node attributes and graph structure, but the node attributes are generally represented as numerical feature vectors rather than natural-language texts.

Text-attributed graphs (TAGs) associate nodes with natural-language descriptions and provide textual semantics beyond conventional node attributes. A comprehensive study of TAGs systematically evaluates language models, graph neural networks, and their combinations, highlighting the importance of jointly modeling textual semantics and graph topology (Yan et al., 2023). Recent TAG learning methods therefore pay increasing attention to the interaction between textual and structural information. SKETCH decouples node aggregation from graph convolution and introduces semantic and structural aggregation for TAG representation learning (Zhou et al., 2025). More directly related to clustering, GCLR selectively queries an LLM for feedback and uses the feedback to refine GNN-based graph clustering (Trivedi et al., 2024). MARK further uses multiple LLM-based agents to infer cluster semantics and provide ranking guidance for uncertain nodes in text-attributed graph clustering (Fu et al., 2025b). These studies show that raw node texts can provide useful semantic information for graph clustering. However, they mainly assume centralized access to node texts and graph structure, where clustering and semantic reasoning are performed within a single learning system. FedN-TAGC instead considers text-attributed subgraphs distributed across multiple clients, where local clusters must be aligned without sharing raw graph data and textual and structural views may lead to different cluster assignments.

A.2 Federated Node-Level Graph Clustering

Node-level federated graph learning enables multiple clients to collaboratively learn from distributed subgraphs while keeping local graph data private. Unlike conventional federated learning with independent samples, graph partitioning can remove cross-subgraph links and leave each client with incomplete local neighborhoods. FedSage+ addresses missing neighborhood information by generating missing neighbors for local subgraphs (Zhang et al., 2021). Subsequent studies further consider structural and distributional heterogeneity across clients. FedTAD addresses node and topology variation and transfers reliable class-wise knowledge through topology-aware data-free knowledge distillation (Zhu et al., 2024). FedStruct captures deep structural dependencies across interconnected subgraphs using explicit global graph structure information (Aliakbari et al., 2025). FedLoG constructs global synthetic data from reliable class representations and structural information to improve local generalization to unseen nodes (Kim et al., 2025). FedATH addresses topology heterogeneity from a causal perspective by separating local graphs into causal and biased subgraphs (Fu et al., 2025a), while FedSPA further studies homophily heterogeneity across clients (Tan et al., 2025). These methods mainly focus on supervised or semi-supervised node prediction rather than node clustering.

More closely related to our setting, recent studies have begun to investigate federated node-level graph clustering without node labels. FedNCN formulates federated node-level graph clustering and uses clustering signals to mend missing cross-subgraph links (Liu et al., 2025). It further learns consensus prototypes on the server to support local node clustering. IFedGC extends this setting to graphs with incomplete node attributes by constructing a global causal structure and aggregating high-confidence cluster centroids for cross-client attribute completion (Liu et al., 2026). Both methods advance federated node-level graph clustering, but they mainly operate on graph topology and vector-valued

node attributes. Neither method explicitly models raw node texts or the case where textual and structural views lead to different cluster assignments.

Recent studies have also started to incorporate textual information into federated graph learning. LLM4FGL uses large language models to generate missing neighbors and infer links for local text-attributed graphs (Yan et al., 2025). DANCE introduces graph condensation into federated text-attributed graph learning to reduce the token and computation costs of LLM-based text processing (Chen et al., 2026). Shi et al. study multimodal adversarial threats in federated TAG learning and propose STRUM to improve robustness against attacks on textual and structural information (Shi et al., 2026). These studies extend federated graph learning to text-attributed graphs, but they mainly focus on supervised node learning, data construction, efficiency, or robustness rather than unsupervised node clustering. FedN-TAGC instead studies federated node-level clustering on text-attributed graphs. It uses local cluster descriptions to support cross-client cluster alignment while explicitly accounting for structure-text conflicts at clients.

B NOTATIONS

Table 3 summarizes the main symbols used in FedN-TAGC. The notation separates private client-side observations, local reliability diagnosis, semantic uploads, global topics, and optimization terms.

Table 3: Summary of key notations in FedN-TAGC.

Symbol	Description	Symbol	Description
M	Number of clients	K	Number of latent clusters/topics
$\mathcal{G}_m$	Local TAG on client m	$\mathcal{V}_m, \mathcal{E}_m$	Local nodes and edges
n_m	Number of local nodes	$\mathcal{T}_m$	Raw node texts on client m
$\mathbf{X}_m$	Text-derived node features	$\mathbf{A}_m$	Normalized local adjacency
$\mathbf{I}_m$	Identity adjacency	f_{θ_m}	Local graph encoder
$\mathbf{Z}_m$	Graph-propagated clustering representation	$\mathbf{Z}_m^t$	Text-view representation
$\mathbf{Z}_m^s$	Structure-view representation	$\mathbf{C}_m$	Local clustering centers
$y_{m,i}$	Local assignment of node i	$a_{m,i}^t$	Nearest text-view cluster
$a_{m,i}^s$	Nearest structure-view cluster	$r_{m,i}$	Cross-view disagreement score
$\tau_{m,k}$	Cluster-wise high-tail threshold	$b_{m,i}$	Calibration-candidate indicator
$\mathcal{S}_{m,k}$	Stable evidence set	$\mathcal{B}_m$	Calibration candidate set
$s_{m,k}$	Local cluster summary	$u_{m,k}$	Uploaded summary tuple
$n_{m,k}$	Stable evidence count	$\mathcal{H}$	Global topic set
h_g	Global topic description	$\pi_m(k)$	Local-to-global topic mapping
$\phi(\cdot)$	Topic text encoder	$\mathcal{C}_m[i]$	Calibrated topic assignment
$\mathcal{L}_{\text{rec}}$	Reconstruction loss	$\mathcal{L}_{\text{clu}}$	Deep clustering loss
$\mathcal{L}_{\text{st}}$	Stable consistency loss	$\mathcal{L}_{\text{gsc}}$	Global semantic consensus loss
λ_{st}	Stable consistency weight	λ_{gsc}	Semantic consensus weight

C ADDITIONAL ANALYSIS OF FEDN-TAGC

FedN-TAGC uses neural encoders, K-means assignments, and LLM-based semantic operations, so the full training process is non-convex and discrete. We do not claim global optimality. This section clarifies the method-specific properties that motivate the design: reliability-aware evidence selection, many-to-one topic alignment, and compact communication.

C.1 RELIABILITY-AWARE EVIDENCE SELECTION

For a local cluster k on client m, FedN-TAGC summarizes only the stable evidence set

$$\mathcal{S}_{m,k} = \{i \mid y_{m,i} = k,\ b_{m,i} = 0\}. \tag{13}$$

The indicator $b_{m,i}$ depends on both cross-view assignment disagreement and a cluster-wise high-tail threshold. This has two effects. First, a node is not treated as ambiguous merely because it has a large

distance to its local center; its text-view and structure-view assignments must also disagree. Second, the threshold is computed within each local cluster, so large or diffuse clusters are not judged by the same absolute scale as compact clusters.

This design reduces semantic contamination in local summaries. If the server receives a summary built from all nodes in a local cluster, a small number of structure-text conflict nodes can introduce a different topic into the description. By excluding $\mathcal{B}_m$ before summary generation, the upload $u_{m,k}$ is more likely to describe the dominant stable evidence of the local cluster. The ambiguous nodes are not discarded; they are delayed until global topics are available.

C.2 Many-to-One Topic Alignment

FedN-TAGC does not require a permutation between a client's K local clusters and the K global topics. The server estimates a mapping $\pi_m : \{1, \dots, K\} \to \{1, \dots, K\}$, and the mapping can be many-to-one. This is important under non-IID client partitions. A client may not contain nodes from every global topic, and two local clusters may correspond to subregions of the same global topic. Forcing a one-to-one assignment would introduce artificial matches and corrupt the broadcast topic descriptions.

The global topic set $\mathcal{H}$ therefore acts as a semantic coordinate system rather than a balanced matching constraint. Local cluster IDs are temporary client-side indices; global topic IDs are the cross-client semantic indices used for final aligned assignments and for resolving calibration candidates.

C.3 Communication and Server Visibility

The semantic path uploads one natural-language summary per non-empty local cluster, plus a stable evidence count. If L_s denotes the average token length of a summary, the semantic upload cost per round is

$$\mathcal{O}(MKL_s). \tag{14}$$

This cost does not grow with the number of local nodes except through the preset number of clusters. In the optional FedNCN-style proxy path, clients upload transformed and noised proxy features selected near local cluster centers and a proxy graph manifest. The saved server-side audit files contain summaries, topic broadcasts, mappings, and proxy shapes/statistics; they do not save raw texts, raw SBERT features, original local edges, node identities, dense node embeddings, disagreement scores, or labels.

D Details of Benchmark Datasets

We evaluate FedN-TAGC on five text-attributed graph benchmarks. Nodes are documents or web pages, edges are citation or hyperlink relations, and node attributes are derived from raw texts. Labels are used only after training to compute clustering metrics.

Table 4: Statistics of text-attributed graph benchmark datasets. "Edges" reports undirected edges when the public benchmark gives an undirected graph; for the SBERT files, the edge-index entry count can be twice the undirected count due to bidirectional storage.

Dataset	Nodes	Edges	Classes	Text feature
Cora	2,708	5,429	7	SBERT-384
PubMed	19,717	44,335	3	SBERT-384
WikiCS	11,701	216,123	10	Text-derived
CiteSeer	3,186	4,225	6	SBERT-384
Wisconsin	251	515	5	Text-derived

For Cora, CiteSeer, and PubMed, we use the public SBERT-preprocessed files from the MARK text-attributed graph benchmark source. The downloaded files contain 384-dimensional SBERT features and raw node texts. In this preprocessing version, CiteSeer contains 3,186 nodes; this differs from the commonly reported 3,327-node citation-network variant and is the version used in our

experiments. The Cora and PubMed SBERT files contain 2,708 and 19,717 nodes, respectively. WikiCS and Wisconsin are included as additional text-attributed graph benchmarks in the main comparison to test larger web-page graphs and small heterophilous web-page graphs.

Before training, features are ℓ_2 normalized on each client. The graph is partitioned into induced local subgraphs; each client receives only its local nodes, local induced edges, node texts, and text features. Cross-client edges are not given to clients during local message passing. Ground-truth labels are not used for partitioning, representation learning, conflict detection, summary generation, topic discovery, alignment, calibration, or server aggregation.

E Details of Federated Settings

Each benchmark graph is partitioned into $M \in \{5, 10, 20\}$ clients. The main experiments report all three client scales. We use topology-aware partitioning: when METIS is available, the graph is partitioned by METIS; otherwise we use a recursive Kernighan–Lin bisection fallback. Both partitioners use only graph topology and never use node labels or node texts. After partitioning, client m receives the induced subgraph on its assigned node set.

This protocol creates a realistic non-IID node-level setting. Because clients own connected or approximately connected graph regions, local class/topic proportions can differ substantially across clients. Some clients may miss a global topic entirely, while others may contain multiple local clusters from the same global topic. FedN-TAGC therefore allows many-to-one local-to-global topic alignment instead of enforcing a permutation between each client's local clusters and the global topic set.

Let $\mathrm{owner}(i)$ denote the client that owns node i. Edges (i, j) with $\mathrm{owner}(i) \neq \mathrm{owner}(j)$ are cut from the local induced subgraphs and are not used in client-side message passing. The server also does not receive these original cross-client edges. The partition record stores client sizes, total edges, cut edges, and cut ratio for audit purposes, but these statistics are not used as labels or semantic supervision.

For each dataset and client scale, the cluster number K is set to the number of ground-truth classes only as a standard clustering-evaluation convention. The class labels themselves remain hidden from training and are used only by the post-hoc ACC, NMI, ARI, and F1 calculations.

F Details of Baseline Methods

We compare FedN-TAGC with representative federated graph clustering and federated graph learning baselines under the same dataset, client scale, and evaluation protocol. For methods originally designed for supervised learning, we remove training-time label supervision and evaluate the learned representations or adapted clustering outputs with post-hoc clustering metrics.

FedNCN. FedNCN is a federated node-level clustering method with cross-subgraph link mending (Liu et al., 2025). It is the closest topology-driven baseline. In our comparison, it uses the same client partitions and does not access node labels during training.

IFedGC. IFedGC studies incomplete federated graph clustering and uses causally-aware attribute completion to improve clustering under incomplete attributes (Liu et al., 2026). We include it as a recent federated graph clustering baseline and run its unsupervised clustering branch under the same client settings.

FedPUB, FedTAD, FedGTA, and FedIIH. These methods are federated graph learning baselines originally developed for supervised or task-specific graph learning. FedTAD uses topology-aware data-free knowledge distillation for subgraph federated learning (Zhu et al., 2024), and FedGTA uses topology-aware averaging (Li et al., 2024). In our unsupervised protocol, label-based objectives and classifier heads are removed or disabled, and the resulting representations are clustered for evaluation. We denote these adapted versions with the suffix "-U" in the main table.

Federated aggregation strategies. For the aggregation ablation, we keep the same local clustering backbone and replace the server-side collaboration strategy with FedAvg, FedProx, or FedPer. FedAvg performs standard parameter averaging (McMahan et al., 2017); FedProx adds a proximal regularizer

to reduce client drift; FedPer keeps a personalized branch while aggregating shared parameters. These ablations test whether generic federated optimization alone can explain the gains of FedN-TAGC.

All baselines are evaluated with the same post-hoc metrics. Ground-truth labels are not used by any method during representation learning, federated communication, cluster alignment, or model aggregation.

G More Implementation Details

Backbone. Each client uses an SDCN/FedNCN-style graph autoencoder. The encoder dimensions are 500-500-2000-10, where the final dimension $n_z = 10$ is the latent clustering representation. The decoder mirrors this structure with 2000-500-500 hidden dimensions. The local graph input is the row-normalized induced adjacency matrix with self-loops. The text-view diagnostic pass uses an identity adjacency matrix so that no neighborhood propagation is applied.

Local and global training. Unless otherwise specified, the final runner uses $R = 3$ communication rounds. Each client performs 10 local epochs per round with Adam and learning rate 10^{-3}. The server-side proxy model is trained for 10 epochs per round with Adam and learning rate 10^{-3}. K-means uses 20 initializations. The default device is selected automatically, using CUDA when available and CPU otherwise.

Reliability diagnosis. After local training, each client runs K-means on the graph-propagated clustering representation. The resulting centers are used to compute nearest-cluster assignments for both the text-only diagnostic representation and the graph-propagated representation. A node becomes a calibration candidate only if the two view assignments disagree and its disagreement score is above the cluster-wise 0.9 quantile. Already fixed calibration results are not selected again in later rounds.

Semantic summarization and calibration. For each non-empty local cluster, the client selects central and diverse stable texts. At most seven representative documents are included in a summary prompt, and each document is truncated to 1,000 characters before prompting. The prompt asks for a compact 40–70 word summary of the main research subject. We use GPT-5 through a Responses-API-compatible endpoint for local summary generation, global topic discovery, topic refinement, local-to-global alignment, and calibration of ambiguous nodes. The API is called with low reasoning effort, JSON-only instructions, no explicit temperature override, and operation-specific output limits of 1,200–4,000 tokens; all responses are validated against the expected JSON schema before use. Representative texts are sent only to the configured LLM endpoint for semantic processing; the federated server and saved audit artifacts receive only generated summaries, topic broadcasts, mappings, and proxy manifests. We therefore treat the LLM endpoint as part of the trusted execution boundary rather than as a formal privacy mechanism. The server groups local summaries into exactly K global topics in the first round, audits the partition, and then keeps topic IDs fixed in later rounds. For calibration candidates, the LLM receives only the candidate text, truncated to 1,800 characters, and the global topic descriptions; it does not receive the current local cluster guess, node embedding, graph edges, neighbor labels, disagreement score, or ground-truth label. Lexical and no-resolution modes are retained for ablations.

Proxy model feedback. In addition to the semantic path, the implementation can upload transformed and slightly noised proxy features selected near local cluster centers. The default selection fraction is 0.1 of the average client size per cluster, with five proxy-MLP epochs and Gaussian noise scale 0.01. The server builds a proxy graph from proxy-feature similarity and local degree statistics, optionally applies N-cut pruning, trains a global proxy model, and broadcasts compatible model parameters and the cluster layer back to clients. This proxy path is an experimental representation-feedback mechanism and should not be interpreted as a formal privacy guarantee.

Audit and reproducibility. The saved audit files include the experiment configuration, source snapshot, server-received summaries, topic broadcasts, local-to-global mappings, proxy manifest shapes/statistics, and private evaluation arrays. The saved server-received directory omits raw texts, raw SBERT features, original graph edges, node IDs, dense node embeddings, disagreement scores,

and labels. The review package provides a hash-only integrity check and an offline result audit; the full audit requires a working PyTorch, NumPy, SciPy, and scikit-learn environment.

H Evaluation Metrics

We evaluate clustering quality with ACC, NMI, ARI, and macro-F1. Labels are never used for training and are used only for post-hoc evaluation.

- **Clustering Accuracy (ACC).** ACC is computed after optimal one-to-one matching between predicted clusters and ground-truth classes:

$$\mathrm{ACC} = \max_{\pi} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(y_i = \pi(\hat{y}_i)\right), \tag{15}$$

  where π is the optimal cluster-to-label permutation.

- **Normalized Mutual Information (NMI).** NMI measures the normalized mutual dependence between the predicted partition $\hat{Y}$ and ground-truth labels Y:

$$\mathrm{NMI} = \frac{I(Y, \hat{Y})}{\sqrt{H(Y)H(\hat{Y})}}. \tag{16}$$

- **Adjusted Rand Index (ARI).** ARI evaluates pairwise agreement between two partitions while correcting for chance:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}. \tag{17}$$

- **Macro-F1.** Macro-F1 averages class-wise F1 scores after the same optimal label matching used for ACC:

$$\mathrm{F1} = \frac{1}{C} \sum_{c=1}^{C} \frac{2\,\mathrm{Precision}_c\,\mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}. \tag{18}$$

All scores in the tables are reported as mean and standard deviation over the specified random seeds and are multiplied by 100.

I Additional Aggregation Ablation Results

Table 5 reports the remaining aggregation-strategy ablation results for Client 10 and Client 20. The comparison uses the same local encoder, clustering objective, and evaluation metrics as the main experiments; the Local row disables cross-client collaboration, and the remaining rows change only the server-side aggregation strategy.

J Additional Experimental Figures

J.1 Additional Module Ablation Results

Figures 6–8 provide the complete module-ablation curves under different client scales. These figures complement the representative main-text results and show that removing both LSRC and GSC is consistently the weakest setting, while the two single-module removals vary across datasets and metrics.

J.2 Additional Convergence Results

Figures 9–11 report the convergence curves for all remaining client-scale settings. The curves preserve the same trend as the main text: performance improves rapidly in the early communication stage, shows small stochastic fluctuations before convergence, and reaches its best value at the final reported round.

Table 5: Aggregation strategy ablation on Client 10 and Client 20. **Bold** and underlined values denote the best and second-best performance, respectively.

Dataset	Strategy	Client 10				Client 20			
		ACC	NMI	ARI	F1	ACC	NMI	ARI	F1
Cora	Local	47.50±4.04	26.47±2.22	20.02±3.70	41.68±3.74	50.35±2.53	23.64±2.95	17.24±3.61	46.11±2.90
	FedAvg	51.22±0.32	0.92±0.12	0.53±0.16	4.93±0.34	49.89±1.00	0.55±0.27	0.28±0.52	0.47±0.77
	FedProx	8.91±0.66	2.49±0.01	0.05±0.01	7.31±0.26	7.57±0.93	1.85±0.32	0.60±0.71	3.78±0.74
	FedPer	8.79±0.26	0.49±0.76	0.26±0.16	3.85±0.22	5.08±0.49	0.30±0.22	0.20±0.65	0.08±0.78
	Ours	**62.92±2.35**	**43.98±0.47**	**40.24±1.30**	**59.04±1.17**	**64.49±0.39**	**42.02±0.73**	**36.71±1.55**	**62.46±0.50**
PubMed	Local	46.88±4.23	0.68±0.16	0.39±0.17	44.17±4.74	42.56±4.34	0.63±0.34	0.62±0.25	41.61±4.36
	FedAvg	54.09±1.60	4.88±0.16	2.74±0.24	26.82±1.49	47.50±0.70	0.34±0.77	0.19±0.48	29.12±0.35
	FedProx	34.85±0.11	4.32±0.65	4.02±0.58	25.88±0.54	35.58±0.42	7.98±0.32	0.70±0.81	25.86±0.27
	FedPer	37.54±0.38	3.83±0.54	0.43±0.01	29.41±0.91	37.92±0.07	3.89±0.56	0.53±0.33	27.18±0.78
	Ours	**61.67±2.59**	**17.97±2.08**	**19.80±5.86**	**60.66±2.69**	**58.12±3.18**	**13.61±0.51**	**14.71±3.39**	**57.38±2.32**
WikiCS	Local	44.63±4.10	23.18±2.85	21.99±2.27	32.53±4.82	40.66±4.58	23.13±2.51	17.64±4.34	28.53±4.81
	FedAvg	47.80±0.08	6.41±0.22	2.94±0.44	13.27±0.01	43.18±0.08	7.16±0.25	4.12±0.38	10.46±0.01
	FedProx	11.53±0.56	0.75±0.46	0.02±0.57	7.23±0.38	3.86±0.78	0.49±0.92	0.32±0.10	2.15±0.29
	FedPer	17.55±0.36	9.24±0.36	7.70±0.01	10.94±0.32	14.77±0.53	9.13±0.21	7.71±0.01	11.96±0.46
	Ours	**59.34±1.88**	**41.41±0.96**	**41.57±0.22**	**48.55±2.56**	**55.59±1.93**	**41.70±0.59**	**37.65±2.17**	**44.30±2.88**
CiteSeer	Local	38.70±3.56	9.46±3.92	6.30±3.93	31.21±2.85	35.89±4.24	6.90±3.39	4.25±3.66	29.66±3.07
	FedAvg	42.14±0.08	6.13±0.81	7.66±0.01	24.25±0.48	39.47±0.42	9.58±0.36	3.35±0.61	12.04±0.01
	FedProx	20.07±0.28	0.85±0.63	0.43±0.31	16.34±1.02	22.66±0.43	1.58±0.27	0.61±0.23	18.21±0.54
	FedPer	29.96±0.82	6.76±0.30	2.91±0.78	11.07±0.01	15.49±0.10	6.35±0.33	4.73±0.18	24.39±0.67
	Ours	**53.12±1.53**	**26.63±1.58**	**25.28±1.71**	**47.69±0.39**	**51.05±1.80**	**25.32±1.20**	**23.73±1.74**	**46.22±0.96**
Wisconsin	Local	51.86±3.56	30.15±2.50	30.69±5.64	33.54±4.90	51.48±2.72	24.36±3.50	29.60±3.15	35.44±3.12
	FedAvg	57.46±0.87	6.19±2.01	0.64±1.67	9.47±1.57	53.95±0.79	9.64±1.41	0.39±0.77	10.16±0.82
	FedProx	10.78±0.26	4.00±1.07	0.44±0.59	9.61±0.30	14.79±1.04	11.81±0.86	0.55±1.78	4.50±0.48
	FedPer	10.41±0.36	5.55±1.58	0.55±1.01	10.19±0.24	13.35±1.23	15.67±0.40	0.29±1.57	1.96±0.62
	Ours	**66.79±1.60**	**49.42±0.68**	**49.39±5.20**	**50.80±3.93**	**66.79±0.53**	**43.63±1.24**	**48.89±1.03**	**51.68±0.69**

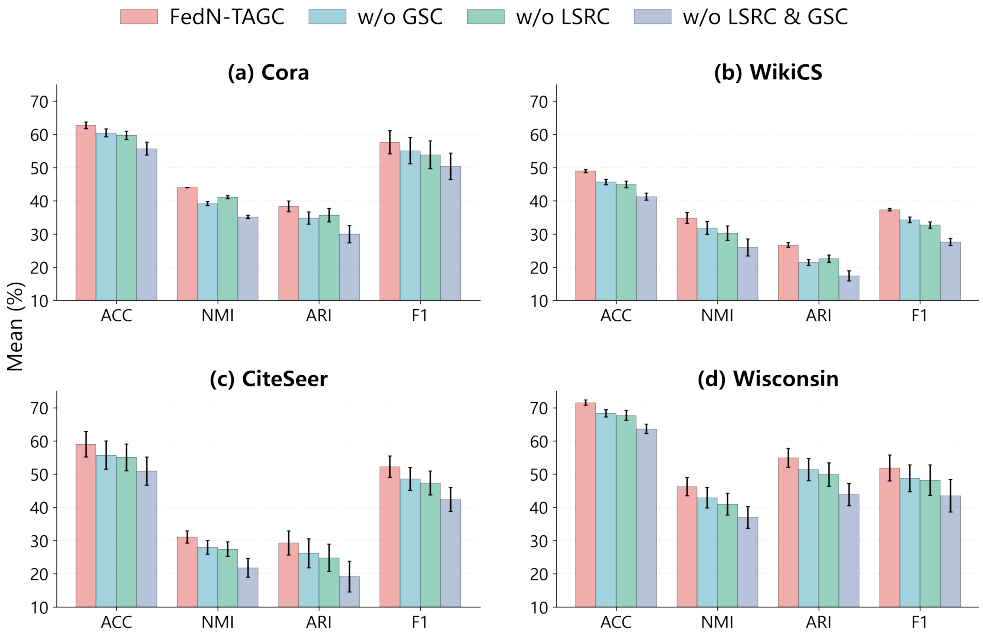

Figure 6: Complete module ablation results under Client 5.

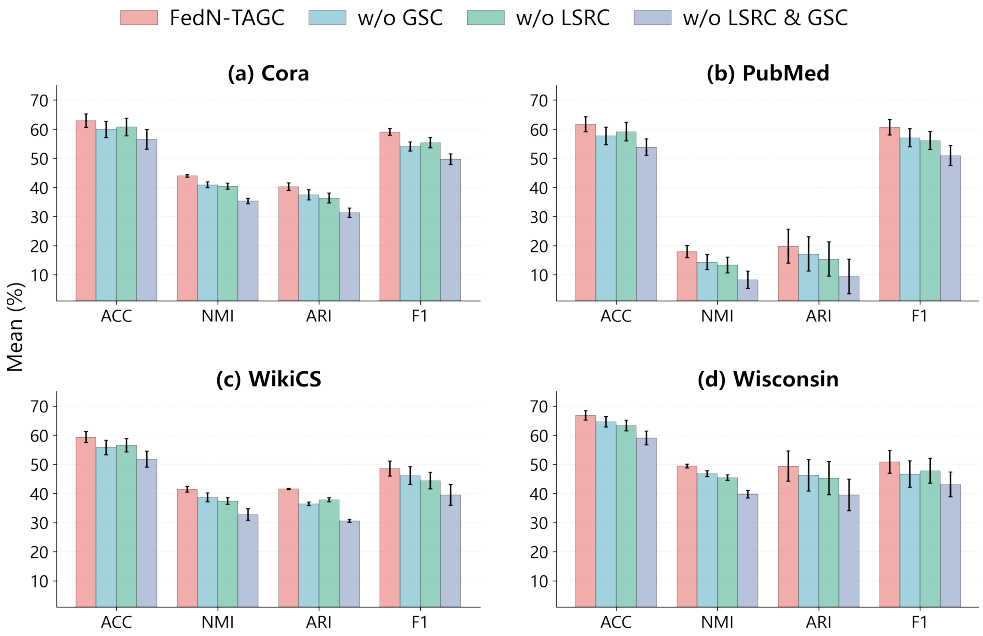

Figure 7: Complete module ablation results under Client 10.

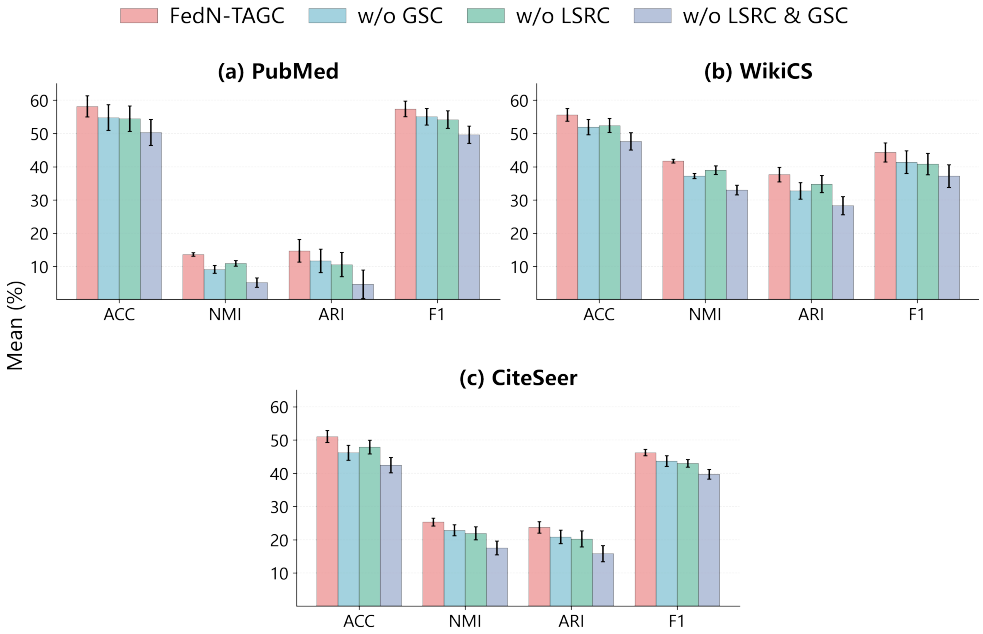

Figure 8: Complete module ablation results under Client 20.

J.3 Additional Hyperparameter Results

Figures 12–14 show the ACC sensitivity to λ_{st} and λ_{gsc} for additional client scales. The balanced setting remains favorable, while nearby settings are usually close, indicating that FedN-TAGC is stable within a reasonable range of loss weights.

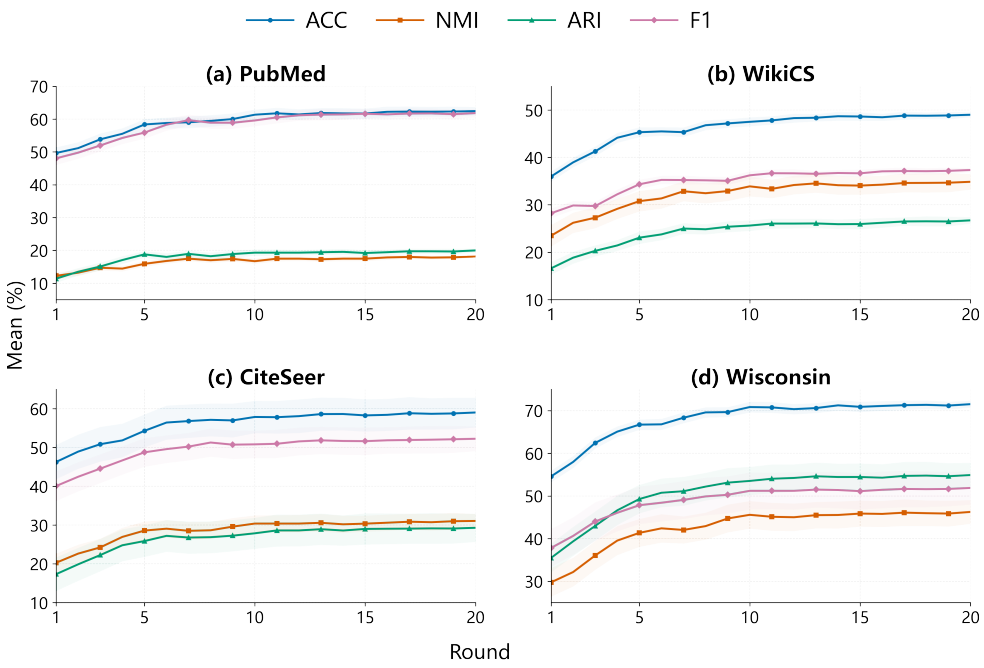

Figure 9: Convergence analysis under Client 5.

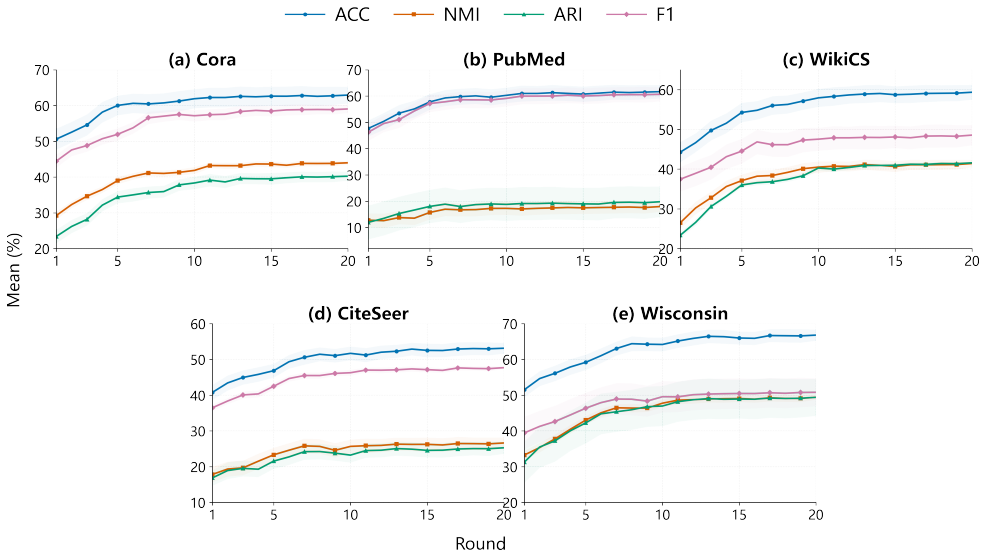

Figure 10: Convergence analysis under Client 10.

K LIMITATIONS

FedN-TAGC focuses on unsupervised federated node-level clustering for text-attributed graphs. Several limitations remain. First, the method relies on the quality of the local text features and graph encoder. If the text encoder does not capture domain semantics or the local induced subgraph is extremely sparse, both the stable evidence selection and local clustering backbone can be affected. Second, the semantic path uses LLM-generated summaries and topic assignments. Although the prompts use strict JSON validation and exclude labels and graph-side private fields, LLM responses can still vary across providers or model versions. Third, the current implementation treats global topic descriptions as semantic anchors rather than providing a formal privacy mechanism. Natural-language

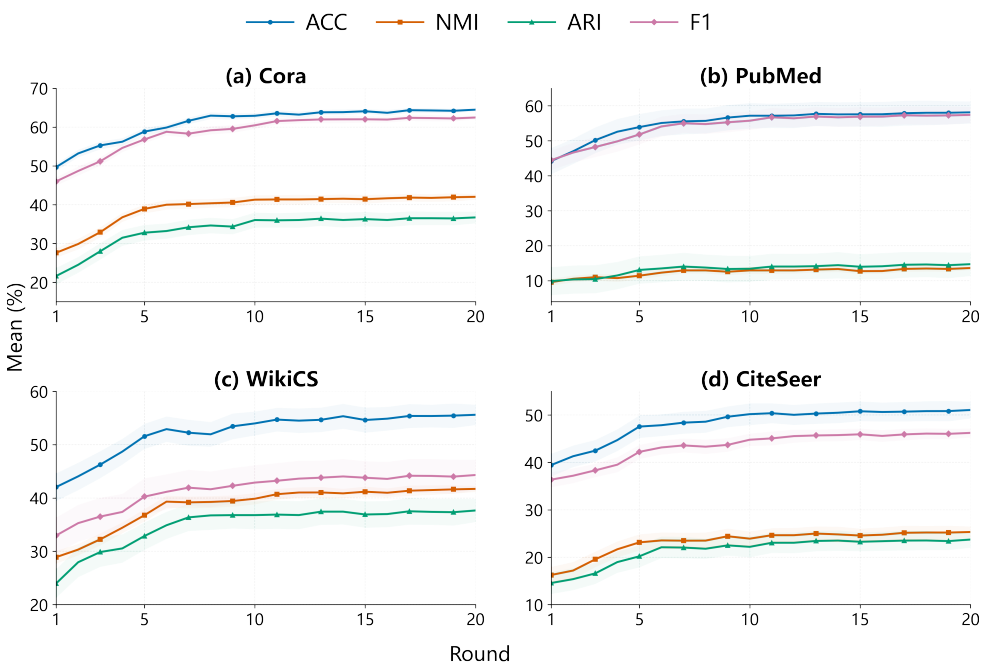

Figure 11: Convergence analysis under Client 20.

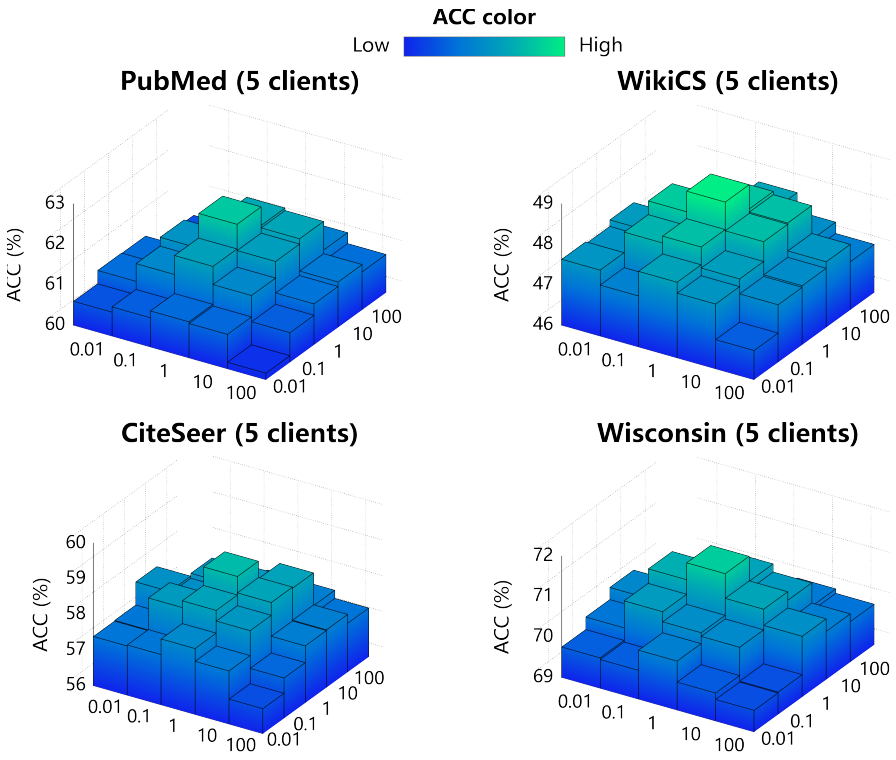

Figure 12: Hyperparameter sensitivity on ACC under Client 5.

summaries may still reveal high-level information about local data distributions, so stronger privacy protections such as differential privacy are left for future work. Finally, the experiments use standard benchmark partitions; deployment on dynamic graphs, streaming texts, or clients with changing topic coverage requires further study.

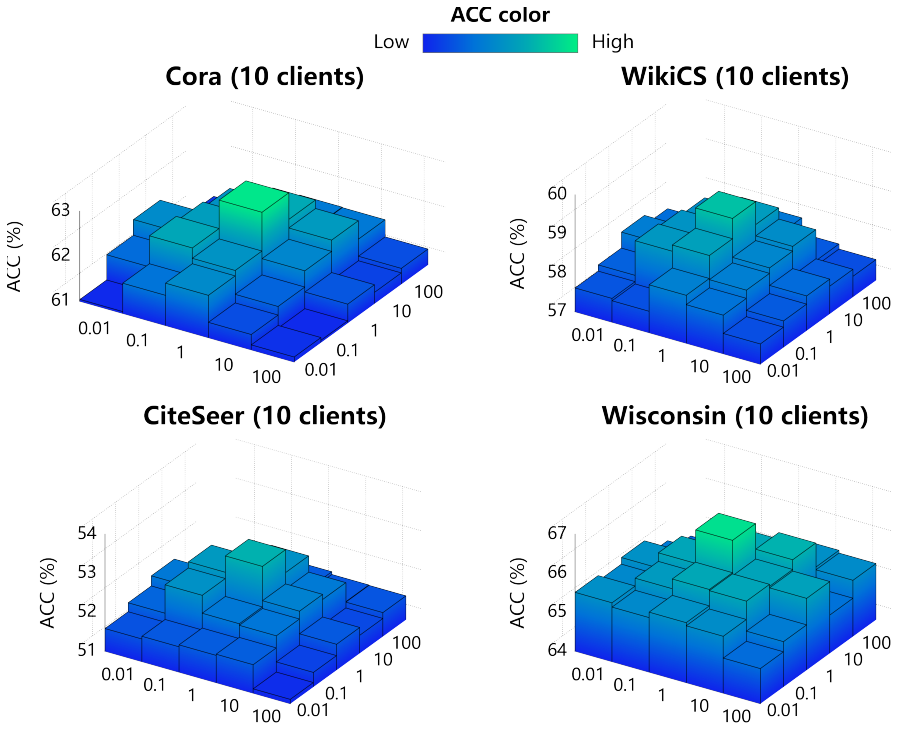

Figure 13: Hyperparameter sensitivity on ACC under Client 10.

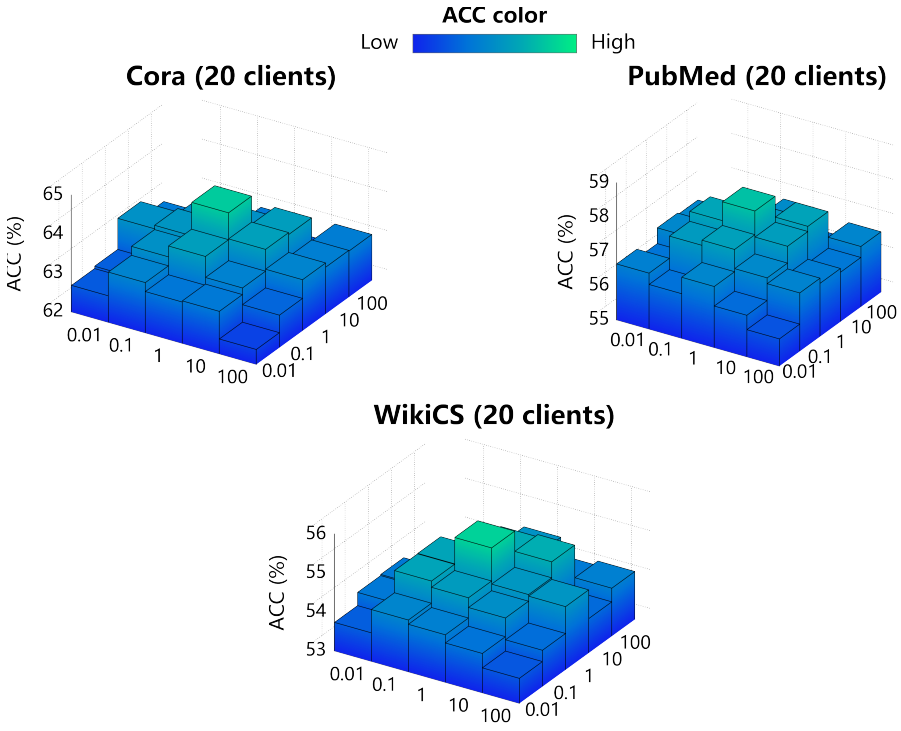

Figure 14: Hyperparameter sensitivity on ACC under Client 20.

L BROADER IMPACTS

This work studies general federated clustering for text-attributed graphs. It can help organizations analyze distributed graph-text data without centralizing raw node texts or graph structures. At the same time, clustering results should not be interpreted as ground-truth categories without domain validation. In high-stakes domains, practitioners should audit the semantic topics, check for spurious structure-text conflicts, and evaluate whether summaries reveal more information than intended before deployment.