
Personalized Subgraph Federated Learning with Differentiable Auxiliary Projections

Wei Zhuo¹, Zhaohuan Zhan², Han Yu¹

¹Nanyang Technological University, ²Shenzhen MSU-BIT University

¹{wei.zhuo, han.yu}@ntu.edu.sg, ²zhan.z@smbu.edu.cn

Abstract

Federated Learning (FL) on graph-structured data typically faces non-IID challenges, particularly in scenarios where each client holds a distinct subgraph sampled from a global graph. In this paper, we introduce **F**ederated learning with **A**uxiliary projections (FedAux), a personalized subgraph FL framework that learns to align, compare, and aggregate heterogeneously distributed local models without sharing raw data or node embeddings. In FedAux, each client jointly trains (i) a local GNN and (ii) a learnable auxiliary projection vector (APV) that differentially projects node embeddings onto a 1D space. A soft-sorting operation followed by a lightweight 1D convolution refines these embeddings in the ordered space, enabling the APV to effectively capture client-specific information. After local training, these APVs serve as compact signatures that the server uses to compute inter-client similarities and perform similarity-weighted parameter mixing, yielding personalized models while preserving cross-client knowledge transfer. Moreover, we provide rigorous theoretical analysis to establish the convergence and rationality of our design. Empirical evaluations across diverse graph benchmarks demonstrate that FedAux substantially outperforms existing baselines in both accuracy and personalization performance. The code is available at <https://github.com/JhuoW/FedAux>.

1 Introduction

Real-world data often manifests as relational structures, ranging from social interactions [30, 46] and financial networks [28, 45] to molecular graphs [35, 44], whose scale and privacy constraints increasingly require training to be carried out in a federated manner [9], whereby multiple clients collaboratively learn a Graph Neural Network (GNN) model without exchanging their raw data. However, applying federated learning to graph-structured data, such as social networks, faces severe challenges due to *non-identically and independently distributed* (non-IID) data across clients. For example, consider a federated learning scenario involving multiple regional social networking platforms, each representing a distinct subgraph of a global social network. Users within each region exhibit unique interaction patterns and distinct interests, resulting in significant heterogeneity in local graph structures and node attributes. This inherent diversity among subgraphs leads to substantial difficulties when attempting to aggregate local GNN models into a unified global model, as traditional FL algorithms [20, 14] typically assume homogeneous data distributions across clients.

To tackle the non-IID challenges inherent in subgraph federated learning, personalized FL [29] has recently emerged as a promising paradigm, which aims to provide client-specific GNN models rather than enforcing a universal global solution. Existing personalized subgraph FL approaches commonly achieve personalization by clustering clients on the server side, necessitating a reliable measure of client similarity without direct access to client-side data. In this work, we impose even stricter privacy constraints: neither raw data nor embeddings are shared, and only model learnable parameters can be exchanged. Although the server could compare clients by directly measuring similarity between their

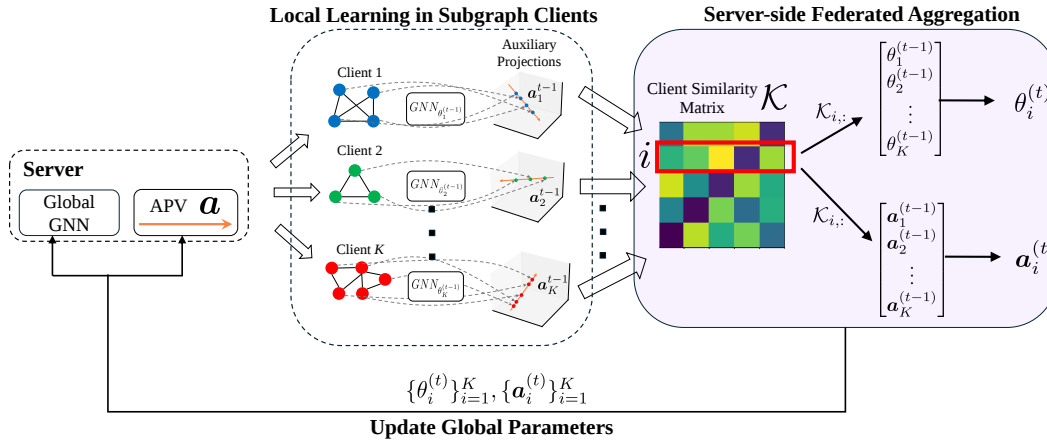


Figure 1: The overall framework of FedAux. Left: The server maintains a global GNN model together with learnable auxiliary projection vectors (APVs) that are broadcast to all clients at the start of each communication round. Middle: Clients jointly optimize the GNN and APV during local training, where the APV projects node embeddings onto a 1D $\mathbf{a}_k$ -space that positions related nodes closer together. Right: After local training, clients transmit their optimized GNN parameters and personalized APVs to the server. The server computes a client similarity matrix by comparing the learned APVs, which captures the heterogeneity across subgraphs without accessing raw data. These similarities determine personalized aggregation weights. At the end of each round, the updated global parameters are broadcast for the next communication round.

parameter matrices uploaded, the high dimensionality of these matrices makes such metrics unreliable under the curse of dimensionality [4]. Recent improvements have proposed measuring similarity by comparing communication-level parameter gradients [35] or generating a common anchor graph on the server as a neutral testbed [3]. Although these strategies mitigate some limitations, they remain largely heuristic and do not explicitly model the heterogeneity inherent in subgraph clients (See extended discussion in Appendix A).

Motivation Our key insight is that a compact, low-dimensional proxy, derived directly from the client’s own model parameters, can faithfully summarize local subgraph characteristics without leaking sensitive node features or embeddings. Such a proxy remains compact enough to avoid the pitfalls of high-dimensional similarity measures, yet expressive enough to reflect meaningful differences between clients. By learning this proxy jointly with the GNN parameters in each client, and using it to guide both local adaptation and server-side aggregation, we obtain a principled, privacy-preserving mechanism for personalization that directly leverages model parameters as a stand-in for subgraph information.

Contribution In this work, we propose FedAux, which employs differentiable auxiliary projections to effectively capture and exploit client-specific heterogeneity for subgraph FL. As illustrated on the left of Fig. 1, the server stores not only a global GNN but also a learnable auxiliary projection vector (APV) that accompanies the model parameters. At the start of the first communication round, the server broadcasts the global GNN and the current APV to all clients. Each client projects its node embeddings onto the APV, which is treated as a one-dimensional latent space. A differentiable soft-sorting operator then orders the projected embeddings by similarity, after which a simulated 1D convolution aggregates the sorted embedding sequence. The aggregated representations drive a supervised loss that simultaneously refines the local GNN and the APV, so that the optimized APV preserves the relational structure of the client subgraph. Upon completing local training, clients send their updated GNN weights and personalized APVs back to the server. Since the APV reveals only the latent space that best preserves local node relationships while concealing the exact position of every node in this space, it acts as a compact privacy-preserving summary of the client subgraph. Then the server computes similarity among the returned APVs to quantify inter-client affinity and yields client-specific aggregation weights. The server uses these weights to combine the incoming parameters, producing a personalized model for each client that respects both shared knowledge and local subgraph idiosyncrasies.

Furthermore, we establish comprehensive and rigorous theoretical analyses that justify the soundness and interoperability of every technique used in FedAux. Extensive federated node classification experiments on six datasets, spanning diverse graph domains and client scales, demonstrate that FedAux achieves better accuracy and stronger personalization than state-of-the-art personalized subgraph FL baselines.

2 Problem Statement: Subgraph Federated Learning

In Federated Learning (FL), multiple clients collaboratively train a global model without exchanging their raw data. In the *subgraph federated learning* setting, each client holds a subgraph of a larger graph. Formally, let a graph $\mathcal{G}$ be partitioned (or subsampled) into K subgraphs $\{G_1, G_2, \dots, G_K\}$ as K clients, where $G_k = (V_k, E_k, X_k, Y_k)$. Here, $V_k = \{v_{k,1}, \dots, v_{k,N_k}\}$ is the set of nodes in the k -th subgraph with $N_k = |V_k|$ nodes, E_k the set of edges among those nodes, $X_k \in \mathbb{R}^{N_k \times d}$ the node features, and Y_k the labels relevant to the learning task. In our FL scenario, each client G_i has access only to its local data (i.e., its subgraph structure, node features, and labels), and there is no sharing of raw data or any node embeddings between clients.

A typical GNN $f_{\theta_k}(G_k)$ parameterized by θ_k is employed to produce node embeddings and ultimately generate predictions on the client G_k . In a standard federated learning setting such as FedAvg [20], one aims to solve the global objective: $\min_{\theta} \sum_{k=1}^K \alpha_k \mathcal{L}_k(\theta)$, subject to the privacy constraint that raw local data G_k never leaves the client side. A common choice is to weight client G_k by $\alpha_k = N_k / (\sum_{j=1}^K N_j)$ or simply $\alpha_k = 1/K$. The iterative procedure proceeds as follows. First, the server initializes $\theta^{(0)}$. At each global communication round $t \in \{1, \dots, T\}$, it sends $\theta^{(t-1)}$ to each client. G_k then updates $\theta^{(t-1)}$ locally by taking a few stochastic gradient steps on $\mathcal{L}_k(\theta)$ to update the parameters $\theta_k \leftarrow \theta_k - \eta \nabla \mathcal{L}$, which produce G_k 's optimal local parameters $\theta_k^{(t)}$. After the t -th local training, all clients' locally updated parameters $\{\theta_1^{(t)}, \dots, \theta_K^{(t)}\}$ are sent back to the server, which aggregates them via a weighted average $\theta^{(t)} = \sum_{k=1}^K \alpha_k \theta_k^{(t)}$. The newly aggregated global parameters $\theta^{(t)}$ are then broadcast back to each client for the next communication round. When the process converges or reaches a designated number of rounds, the final global parameters $\theta^{(T)}$ are taken as the parameters of the GNN model on the server.

3 Methodology

In this section, we introduce the proposed Subgraph Federated Learning with Auxiliary Projections (FedAux) framework, designed to address the heterogeneity across local subgraphs in federated learning. Fig. 1 illustrates an overview of FedAux. Our objective is twofold: 1) Each client locally encodes its subgraph into a one-dimensional space via a learnable auxiliary projection vector APV, and 2) the server then exploits these auxiliary vectors to realize personalized aggregation.

3.1 Client-Side Local Training

Before the first communication round $t = 0$, alongside the GNN model parameterized by $\theta^{(0)}$, the server also maintains a learnable APV $\mathbf{a}^{(0)} \in \mathbb{R}^{d'}$. During each communication round t , the server distributes $(\theta^{(t-1)}, \mathbf{a}^{(t-1)})$ to initialize all clients' local model $\{(\theta_k^{(t-1)}, \mathbf{a}_k^{(t-1)})\}_{k=1}^K \leftarrow (\theta^{(t-1)}, \mathbf{a}^{(t-1)})$. For a client G_k , it runs the local GNN model to optimize the node embeddings:

$$\mathbf{H}_k^{(t-1)} = f_{\theta_k^{(t-1)}}(G_k) = [h_{k,1}^{(t-1)}, h_{k,2}^{(t-1)}, \dots, h_{k,N_k}^{(t-1)}] \in \mathbb{R}^{N_k \times d'}, \quad (1)$$

where d' is the output dimension of node embeddings. Given the local APV $\mathbf{a}_k^{(t-1)}$, we first normalize all node embeddings so that all embeddings are compared on a consistent scale as $\hat{h}_{k,i}^{(t-1)} = h_{k,i}^{(t-1)} / \max_j \|h_{k,j}^{(t-1)}\|$. Then the similarity between node $v_{k,i}$ and $\mathbf{a}_k^{(t-1)}$ is defined as $s_{k,i}^{(t-1)} = \langle \hat{h}_{k,i}^{(t-1)}, \mathbf{a}_k^{(t-1)} \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^{d'}$. Intuitively, $s_{k,i}^{(t-1)}$ can be interpreted as the coordinate of each node $v_{k,i}$ in $\mathbf{a}_k$ -space, which is a 1D line. Since $\mathbf{a}_k^{(t-1)}$ is itself learnable, the client G_k is adaptively refining this space to capture relationships among its node embeddings more effectively.

Next, G_k collects the similarity scores $S_k^{(t-1)} = \{s_{k,1}^{(t-1)}, \dots, s_{k,N_k}^{(t-1)}\}$ and sort them in non-decreasing order. Let π_k be the permutation that orders these scores:

$$s_{k,\pi_k(1)}^{(t-1)} \leq s_{k,\pi_k(2)}^{(t-1)} \leq \dots \leq s_{k,\pi_k(N_k)}^{(t-1)}, \quad (2)$$

where $\pi_k(j)$ represents the node index at rank j . Accordingly, we apply this permutation π_k to the row indices of $\mathbf{H}_k^{(t-1)}$, which yields the sorted embedding matrix $\mathbf{a}_k^{(t-1)}$ as:

$$\tilde{\mathbf{H}}_k^{(t-1)} = \left[\mathbf{H}_k^{(t-1)} \right]_{\pi_k,:} = \left[h_{k,\pi_k(1)}^{(t-1)}, h_{k,\pi_k(2)}^{(t-1)}, \dots, h_{k,\pi_k(N_k)}^{(t-1)} \right], \quad (3)$$

which aligns node embeddings according to their coordinates in the $\mathbf{a}_k$ -space. As shown in Fig. 2, during local training, the node embeddings and the structure of the $\mathbf{a}_k$ -space are jointly optimized so that nodes with stronger relationships are positioned closer along this learned space (in G_k comprising two triangles, nodes within the same triangle should be proximate in the $\mathbf{a}_k$ -space). In other words, the objective of the local model is to adaptively reshape the APV so that the induced node sorting effectively captures the local data information.

Under the semi-supervised setting, the node sorting on APV can be adaptively refined under the guidance of the downstream task. Thinking of each $h_{k,\pi_k(i)}^{(t-1)}$ as

a feature vector in a 1D sequence $\tilde{\mathbf{H}}_k^{(t-1)}$, inspired by [18], we can apply a 1D convolution with a fixed kernel size B over $\tilde{\mathbf{H}}_k^{(t-1)}$ as $\text{Conv1D}([h_{k,\pi_k(1)}^{(t-1)}, h_{k,\pi_k(2)}^{(t-1)}, \dots, h_{k,\pi_k(N_k)}^{(t-1)}])$. More specifically, for each node $v_{\pi_k(i)}$ in the sorted sequence, the convolution can be written as:

$$z_{k,\pi_k(i)}^{(t-1)} = \sum_{\tau=-\lfloor B/2 \rfloor}^{\lfloor B/2 \rfloor} \mathbf{W}_\tau h_{k,\pi_k(i+\tau)}^{(t-1)} + b, \quad (4)$$

where $\mathbf{W}_\tau \in \mathbb{R}^{d' \times d'}$ are learnable convolution kernels for offset τ , and b is a bias term. Convolution around $v_{\pi_k(i)}$ in Eq. (4) amounts to a proximity-based aggregation in the $\mathbf{a}_k$ -space, where each embedding is updated by aggregating information from its neighbors along this learned 1D sorting. Hence, the quality of this sorting significantly impacts the aggregation effectiveness. Intuitively, nodes with stronger semantic relationships or similar labels should appear closer together in this learned sorting. We can formulate the learning objective to explicitly optimize the sorting induced by the APV $\mathbf{a}_k$, ensuring that the resultant sorting facilitates effective aggregation and improves downstream predictive accuracy. Consequently, we formulate the learning objective:

$$(\theta_k^*, \mathbf{a}_k^*, \Phi_k^*) = \underset{\theta_k, \mathbf{a}_k, \Phi_k}{\operatorname{argmin}} \mathcal{L} \left(\text{CLF} \left(\text{Conv1D} \left(\tilde{\mathbf{H}}_k^{(t-1)} \right) \right), Y_k \right), \quad (5)$$

where Φ_k denotes the full set of parameters for Conv1D and the subsequent classifier CLF that maps node embeddings to final logits. Through this objective, we explicitly encourage $\mathbf{a}_k$ -space to yield an optimal node sorting, enabling the convolutional operation to effectively capture and leverage localized, label-informed relationships, thereby the optimized APV $\mathbf{a}_k^{(t)} = \mathbf{a}_k^*$ accurately preserves and encodes the local node relationships specific to each client.

However, $\mathbf{a}_k$ does not directly participate in the loss defined in Eq. (5) in a way that enables standard backpropagation to update it. It is because the role of $\mathbf{a}_k$ is limited to generating similarity scores, which in turn determine the input order to the Conv1D layer. Thus, $\mathbf{a}_k$ only affects the network's output by reordering embeddings, which is a purely indirect pathway that does not produce a gradient signal for $\mathbf{a}_k$ from the downstream loss. [18] attempted to mitigate this by multiplying each node embedding by its similarity score and then sorting. While this modification integrates $\mathbf{a}_k$ into the learning pipeline directly, the hard discrete sort persists, causing the gradient signal that could refine $\mathbf{a}_k$ to be still routed through a non-smooth transformation. Hence $\mathbf{a}_k$ still cannot be fully optimized to reorder embeddings based on loss feedback, leaving the core issue unresolved.

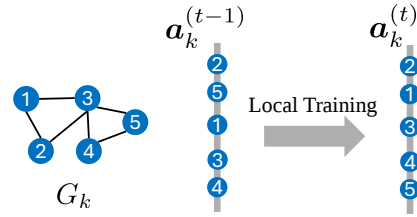


Figure 2: The local training of FedAux aims to map all nodes in G_k to a corresponding $\mathbf{a}_k$ -space, and the optimization objective is to learn the APV $\mathbf{a}_k$, such that the resulting $\mathbf{a}_k$ -space preserves the optimal node sorting.

To eliminate the hard-sorting bottleneck, we propose a *continuous aggregation* scheme over the $\mathbf{a}_k$ -space. Rather than ranking or discretizing these similarity scores, for each node v_i , we define a continuous kernel $\kappa(s_{k,i}^{(t-1)}, s_{k,j}^{(t-1)})$, which could be a simple Gaussian-like function $\mathcal{K}_{ij} = \kappa(s_{k,i}^{(t-1)}, s_{k,j}^{(t-1)}) = \exp\left(-(s_{k,i}^{(t-1)} - s_{k,j}^{(t-1)})^2 / \sigma^2\right)$ with bandwidth $\sigma > 0$, measuring how close v_j is to v_i in the real line spanned by $\{s_k^{(t-1)}\}$. As shown in the right part of Fig. 1, we then obtain an aggregated embedding for each node v_i by a smooth weighted sum of all node embeddings:

$$z_{k,i}^{(t-1)} = \frac{1}{M_i} \sum_{j=1}^{N_k} \kappa(s_{k,i}^{(t-1)}, s_{k,j}^{(t-1)}) h_{k,j}^{(t-1)}, \quad M_i = \sum_{j=1}^{N_k} \kappa(s_{k,i}^{(t-1)}, s_{k,j}^{(t-1)}). \quad (6)$$

Unlike discrete sorting, this continuous aggregator is fully differentiable with respect to $\mathbf{a}_k$, because changes in $\mathbf{a}_k$ smoothly shift each $s_{k,i}^{(t-1)}$ and thus adjust the kernel weights κ . This approach naturally learns to group nodes with similar $s_k^{(t-1)}$ values, emulating the sorted 1D convolution effect, while sidestepping the gradient-blocking issues that arise from hard-sorting steps.

To jointly train the GNN parameters θ_k and the APV $\mathbf{a}_k$, we associate each node v_i with two embeddings: $h_{k,i}^{(t-1)}$ produced by the GNN defined in Eq. (1), and $z_{k,i}^{(t-1)}$ generated via our kernel-based aggregation method as Eq. (6). We then concatenate these embeddings to form v_i 's final embedding $r_{k,i}^{(t-1)} = [h_{k,i}^{(t-1)} \| z_{k,i}^{(t-1)}]$, which is fed into a simple MLP classifier $\text{CLF}(\cdot)$ to produce logits for the cross-entropy loss $\mathcal{L}_k = \frac{1}{N_k} CE(\text{CLF}(\Gamma_k^{(t-1)}), Y_k)$, where $\Gamma^{(t-1)} = [r_{k,i}^{(t-1)}]_{i=1}^{N_k}$.

3.2 Server-Side Federated Aggregation

At the end of local training for communication round t , each client G_k transmits its optimized parameters $(\theta_k^{(t-1)}, \mathbf{a}_k^{(t-1)})$ to the server. In doing so, only these high-level parameters are exchanged, rather than gradients or node embeddings, thereby limiting direct leakage of private subgraph information. Note that $\mathbf{a}_k$ serves as the optimal subspace for capturing node relationships, how individual nodes map into this space (and thus the precise relational details) remains unknown to the server. This design strictly adheres to the fundamental FL principle that *Data stays local; only model updates leave*.

Since the data distributions across clients can be non-IID, the server is expected to personalize the aggregation for each client. Given that each $\mathbf{a}_k^{(t-1)}$ can be a descriptor of how node embeddings in G_k are arranged and structured, the similarity between two clients G_k and G_l can be measured via the cosine similarity of their APVs: $\text{SIM}(\mathbf{a}_k^{(t-1)}, \mathbf{a}_l^{(t-1)}) = \frac{\langle \mathbf{a}_k^{(t-1)}, \mathbf{a}_l^{(t-1)} \rangle}{\|\mathbf{a}_k^{(t-1)}\| \|\mathbf{a}_l^{(t-1)}\|}$. We then convert this similarity into a weight:

$$w_{k,l}^{(t-1)} = \frac{\exp\left(\alpha \text{SIM}\left(\mathbf{a}_k^{(t-1)}, \mathbf{a}_l^{(t-1)}\right)\right)}{\sum_{r=1}^K \exp\left(\alpha \text{SIM}\left(\mathbf{a}_k^{(t-1)}, \mathbf{a}_r^{(t-1)}\right)\right)}, \quad (7)$$

where $\alpha > 0$ is a temperature controlling the sharpness of the weighting distribution. $w_{k,l}^{(t-1)}$ reflects how much client G_k should incorporate the update from G_l . By emphasizing contributions from similar clients (i.e., those with high similarity in their APVs), each client's final model can better handle heterogeneous data while reducing interference from dissimilar clients. Instead of averaging all local parameters into one single global model, the server can compute a personalized aggregation of parameters for each client G_k as:

$$\theta_k^{(t)} = \sum_{l=1}^K w_{k,l}^{(t-1)} \theta_l^{(t-1)}, \quad \mathbf{a}_k^{(t)} = \sum_{l=1}^K w_{k,l}^{(t-1)} \mathbf{a}_l^{(t-1)}. \quad (8)$$

Thus, after the server performs these personalized aggregations for both θ and $\mathbf{a}$, it transmits $(\theta_k^{(t)}, \mathbf{a}_k^{(t)})$ back to client G_k for the $(t+1)$ -th communication round starting point. Appendix B shows the pseudo code of FedAux.

Complexity Analysis For the client side of FedAux, the local GNN embedding generation incurs a complexity of $\mathcal{O}(|E_k|d')$, and the auxiliary projection from embeddings to the APV results in $\mathcal{O}(N_k d')$. Besides, the kernel-based embedding aggregation over the 1D space induced by the APV has complexity $\mathcal{O}(N_k^2 d')$. Consequently, the per-client complexity is $\mathcal{O}(|E_k|d' + N_k^2 d')$. On the server side, computing the client-wise similarity for personalized federated aggregation involves a complexity $\mathcal{O}(K^2 d')$. Therefore, the total complexity of FedAux per communication round is $\mathcal{O}((|E_k| + N_k^2 + K^2)d')$.

3.3 Theoretical Analysis

For notational simplicity, we focus on a single client with N nodes in a given communication round and omit the subscript k and superscript $(t - 1)$. The core of our model is to use a learnable auxiliary projection vector APV $\mathbf{a}$ to capture an optimal node sorting of the local node embeddings and thus serves as a compact summary of the subgraph. However, there is a foundational question that inevitably arises once we replace hard sorting with the smooth kernel aggregator: when the APV $\mathbf{a}$ is learned via back-propagation, *what does it actually learn? Does it encode an arbitrary nuisance direction, or does it converge to a geometrically meaningful axis that faithfully summarizes the local subgraph?* To answer these questions, we analyze the fidelity of the APV with the following theorem.

Theorem 3.1 (Fidelity of the APV $\mathbf{a}$). *Let $\mathbf{C} := \frac{1}{N} \sum_{i=1}^N h_i h_i^\top$ be the empirical covariance of node embeddings in the current client with size N . The gradient of the local loss $\mathcal{L}$ w.r.t. the APV $\mathbf{a}$ satisfies:*

$$\nabla_{\mathbf{a}} \mathcal{L} = -\frac{2}{\sigma^2} \mathbf{C} \mathbf{a} + \mathcal{R}(\sigma), \quad (9)$$

where the remainder term obeys $\|\mathcal{R}(\sigma)\| = \mathcal{O}(\sigma^0)$ as $\sigma \rightarrow 0^+$. Define $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ as the unit Euclidean sphere embedded in $\mathbb{R}^d$, then the gradient descent on $\mathcal{L}$ with unit-norm re-normalization reproduces Oja learning rule [22]:

$$\mathbf{a} \leftarrow \Pi_{\mathbb{S}^{d-1}}(\mathbf{a} - \eta \mathbf{C} \mathbf{a}), \quad (10)$$

whose unique stable fixed points are the eigenvectors of $\mathbf{C}$, and the global attractor is the principal eigenvector (largest eigenvalue).

The proof and more discussions are provided in Appendix C.1. It guarantees that, once the kernel aggregator makes $\mathbf{a}$ differentiable, ordinary back-propagation forces APV to align with the direction along with the node embeddings in that client vary the most. Equivalently, the APV $\mathbf{a}$ is not an arbitrary trainable knob but a statistically optimal, variance-maximizing summary of local embeddings. Thus, **the APV $\mathbf{a}$ is provably the first principal component of the local embeddings.**

In Section 3.1, we propose a continuous kernel aggregator to replace the hard sort-then-Conv1D pipeline used in earlier work [18]. To justify that replacement, the following theorem rigorously shows that the new smooth operator degenerates to the old one in the appropriate limit.

Theorem 3.2 (Sorting limit and equivalence to Conv1D). *Let z_i be the kernel-smoothed embeddings, and gather them in score order $\tilde{\mathbf{Z}} = [z_{\pi(1)}, \dots, z_{\pi(N)}] \in \mathbb{R}^{N \times d'}$. The original sorted embeddings $\tilde{\mathbf{H}} = [h_{\pi(1)}, \dots, h_{\pi(N)}]$ is defined in Eq. (3). Let $\mathbf{W} \in \mathbb{R}^{B \times d'}$ be an arbitrary fixed Conv1D kernel with zero padding, and denote $\text{Conv}_{\mathbf{W}}(\mathcal{X})_t = \sum_{\tau=1}^B \mathbf{W}_{\tau} \mathcal{X}_{t+\tau-\lceil B/2 \rceil}$, for any sequence $\mathcal{X} \in \mathbb{R}^{N \times d'}$. Then we have:*

$$\lim_{\sigma \rightarrow 0^+} \left\| \text{Conv}_{\mathbf{W}}(\tilde{\mathbf{Z}}) - \text{Conv}_{\mathbf{W}}(\tilde{\mathbf{H}}) \right\|_F = 0, \quad (11)$$

where $\|\cdot\|_F$ is the Frobenius norm.

The proof is in Appendix C.2. Theorem 3.2 indicates that the kernel aggregation followed by Conv1D converges to hard-sorting followed by Conv1D as the bandwidth σ tends to 0. Hence, the two architectures have identical expressive power up to an arbitrarily small error for sufficiently small σ . Although the limit $\sigma \rightarrow 0^+$ recovers discrete sorting, a larger σ performs a soft neighborhood pooling that can act as a learnable regularizer against over-fitting noisy local orderings. In practice, we find $\sigma = 1$ to be effective across all datasets.

Next, we present a theoretical analysis of FedAux's convergence rate, which guarantees that it cannot diverge in expectation. Since this analysis focuses on the global model, we use the subscript $\cdot_k$ to denote the client index.

Table 1: Federated node classification results. The reported results are the mean and standard deviation over three different runs. Best performance is highlighted in **bold**.

Methods	Cora			CiteSeer			Pubmed		
	5 Clients	10 Clients	20 Clients	5 Clients	10 Clients	20 Clients	5 Clients	10 Clients	20 Clients
Local	81.30 $\pm$ 0.21	79.94 $\pm$ 0.24	80.30 $\pm$ 0.25	69.02 $\pm$ 0.05	67.82 $\pm$ 0.13	65.98 $\pm$ 0.17	84.04 $\pm$ 0.18	82.81 $\pm$ 0.39	82.65 $\pm$ 0.03
FedAvg	74.45 $\pm$ 5.64	69.19 $\pm$ 0.67	69.50 $\pm$ 3.58	71.06 $\pm$ 0.60	63.61 $\pm$ 3.59	64.68 $\pm$ 1.83	79.40 $\pm$ 0.11	82.71 $\pm$ 0.29	80.97 $\pm$ 0.26
FedProx	72.03 $\pm$ 4.56	60.18 $\pm$ 7.04	48.22 $\pm$ 6.81	71.73 $\pm$ 1.11	63.33 $\pm$ 3.25	64.85 $\pm$ 1.35	79.45 $\pm$ 0.25	82.55 $\pm$ 0.24	80.50 $\pm$ 0.25
FedPer	81.68 $\pm$ 0.40	79.35 $\pm$ 0.04	78.01 $\pm$ 0.32	70.41 $\pm$ 0.32	70.53 $\pm$ 0.28	66.64 $\pm$ 0.27	85.80 $\pm$ 0.21	84.20 $\pm$ 0.28	84.72 $\pm$ 0.31
GCFL	81.47 $\pm$ 0.65	78.66 $\pm$ 0.27	79.21 $\pm$ 0.70	70.34 $\pm$ 0.57	69.01 $\pm$ 0.12	66.33 $\pm$ 0.05	85.14 $\pm$ 0.33	84.18 $\pm$ 0.19	83.94 $\pm$ 0.36
FedGNN	81.51 $\pm$ 0.68	70.12 $\pm$ 0.99	70.10 $\pm$ 3.52	69.06 $\pm$ 0.92	55.52 $\pm$ 3.17	52.23 $\pm$ 6.00	79.52 $\pm$ 0.23	83.25 $\pm$ 0.45	81.61 $\pm$ 0.59
FedGTA	71.26 $\pm$ 2.93	68.33 $\pm$ 1.27	69.24 $\pm$ 0.91	69.39 $\pm$ 0.75	67.34 $\pm$ 1.08	65.29 $\pm$ 1.92	78.47 $\pm$ 0.25	82.79 $\pm$ 0.20	81.92 $\pm$ 0.60
FedSage+	72.97 $\pm$ 5.94	69.05 $\pm$ 1.59	57.97 $\pm$ 12.6	70.74 $\pm$ 0.69	65.63 $\pm$ 3.10	65.46 $\pm$ 0.74	79.57 $\pm$ 0.24	82.62 $\pm$ 0.31	80.82 $\pm$ 0.25
FED-PUB	83.72 $\pm$ 0.18	81.45 $\pm$ 0.12	81.10 $\pm$ 0.64	72.40 $\pm$ 0.26	71.83 $\pm$ 0.61	66.89 $\pm$ 0.14	86.81 $\pm$ 0.12	86.09 $\pm$ 0.17	84.66 $\pm$ 0.54
FedAux	84.57$\pm$0.39	82.05$\pm$0.71	81.60$\pm$0.64	72.99$\pm$0.82	73.16$\pm$0.29	68.10$\pm$0.35	88.10$\pm$0.16	86.43$\pm$0.20	84.87$\pm$0.42

Methods	Amazon-Computer			Amazon-Photo			ogbn-arxiv		
	5 Clients	10 Clients	20 Clients	5 Clients	10 Clients	20 Clients	5 Clients	10 Clients	20 Clients
Local	89.22 $\pm$ 0.13	88.91 $\pm$ 0.17	89.52 $\pm$ 0.20	91.67 $\pm$ 0.09	91.80 $\pm$ 0.02	90.47 $\pm$ 0.15	66.76 $\pm$ 0.07	64.92 $\pm$ 0.09	65.06 $\pm$ 0.05
FedAvg	84.88 $\pm$ 1.96	79.54 $\pm$ 0.23	74.79 $\pm$ 0.24	89.89 $\pm$ 0.83	83.15 $\pm$ 3.71	81.35 $\pm$ 1.04	65.54 $\pm$ 0.07	64.44 $\pm$ 0.10	63.24 $\pm$ 0.13
FedProx	85.25 $\pm$ 1.27	83.81 $\pm$ 1.09	73.05 $\pm$ 1.30	90.38 $\pm$ 0.48	80.92 $\pm$ 4.64	82.32 $\pm$ 0.29	65.21 $\pm$ 0.20	64.37 $\pm$ 0.18	63.03 $\pm$ 0.04
FedPer	89.67 $\pm$ 0.34	89.73 $\pm$ 0.04	87.86 $\pm$ 0.43	91.44 $\pm$ 0.37	91.76 $\pm$ 0.23	90.59 $\pm$ 0.06	66.87 $\pm$ 0.05	64.99 $\pm$ 0.18	64.66 $\pm$ 0.11
GCFL	89.07 $\pm$ 0.91	90.03 $\pm$ 0.16	89.08$\pm$0.25	91.99 $\pm$ 0.29	92.06 $\pm$ 0.25	90.79 $\pm$ 0.17	66.80 $\pm$ 0.12	65.09 $\pm$ 0.08	65.08 $\pm$ 0.04
FedGNN	88.08 $\pm$ 0.15	88.18 $\pm$ 0.41	83.16 $\pm$ 0.13	90.25 $\pm$ 0.70	87.12 $\pm$ 2.01	81.00 $\pm$ 4.48	65.47 $\pm$ 0.22	64.21 $\pm$ 0.32	63.80 $\pm$ 0.05
FedGTA	85.06 $\pm$ 0.82	84.27 $\pm$ 0.71	79.46 $\pm$ 0.28	89.70 $\pm$ 0.67	76.53 $\pm$ 3.21	82.02 $\pm$ 0.78	65.42 $\pm$ 0.09	64.22 $\pm$ 0.08	63.75 $\pm$ 0.18
FedSage+	85.04 $\pm$ 0.61	80.50 $\pm$ 1.30	70.42 $\pm$ 0.85	90.77 $\pm$ 0.44	76.81 $\pm$ 8.24	80.58 $\pm$ 1.15	65.69 $\pm$ 0.09	64.52 $\pm$ 0.14	63.31 $\pm$ 0.20
FED-PUB	90.25 $\pm$ 0.07	89.73 $\pm$ 0.16	88.20 $\pm$ 0.18	93.20 $\pm$ 0.15	92.46$\pm$0.19	90.59 $\pm$ 0.35	67.62 $\pm$ 0.11	66.35 $\pm$ 0.16	63.90 $\pm$ 0.27
FedAux	90.38$\pm$0.08	89.92$\pm$0.15	88.35 $\pm$ 0.96	93.37$\pm$0.26	92.30 $\pm$ 0.29	90.91$\pm$0.60	68.83$\pm$0.15	68.50$\pm$0.27	65.52$\pm$0.10

Theorem 3.3 (Global linear convergence). *Let $\Psi_k^{(t)} = (\theta_k^{(t)}, \mathbf{a}_k^{(t)})$ be the local parameters of client G_k at the communication round t , and $\Psi^{(t)} = [\Psi_1^{(t)}, \dots, \Psi_K^{(t)}]$ collects all local parameters. Assuming 1) every local objective is $\mathcal{L}$ -smooth: $\forall \Psi_k, \Psi'_k : \|\nabla \mathcal{L}_k(\Psi_k) - \nabla \mathcal{L}_k(\Psi'_k)\| \leq \mathcal{L} \|\Psi_k - \Psi'_k\|$; 2) the stochastic gradients are unbiased ($\mathbb{E}[g_k] = \nabla \mathcal{L}_k$) and variance-bounded ($\mathbb{E}[\|g_k - \nabla \mathcal{L}_k\|^2] \leq \zeta^2$), where $g_k := \nabla_{(\Psi_k)} \mathcal{L}_k$ means the local gradients; 3) each local objective satisfies the μ -PL condition [23]; 4) let $\Omega^{(t)} = [w_{kl}^{(t)}]_{k,l} \in \mathbb{R}^{K \times K}$, the spectral gap $\rho := \sup_t \|\Omega^{(t)} - \frac{1}{K} \mathbf{1}\mathbf{1}^\top\|_2 < 1$. Let each client perform Q local updates per round, and the learning rate $0 < \eta \leq \frac{1}{2\mathcal{L}}$. With any initial parameters $\Psi^{(0)} = (\theta^{(0)}, \mathbf{a}^{(0)})$, we have:*

$$\mathbb{E}[\mathcal{L}(\Psi^{(T)}) - \mathcal{L}^*] \leq (1 - \eta\mu)^{QT} (\mathcal{L}(\Psi^{(0)}) - \mathcal{L}^*) + \frac{\eta\mathcal{L}\zeta^2}{2\mu} + \frac{2\eta\mathcal{L}\rho^2}{\mu(1 - \rho)^2}, \quad (12)$$

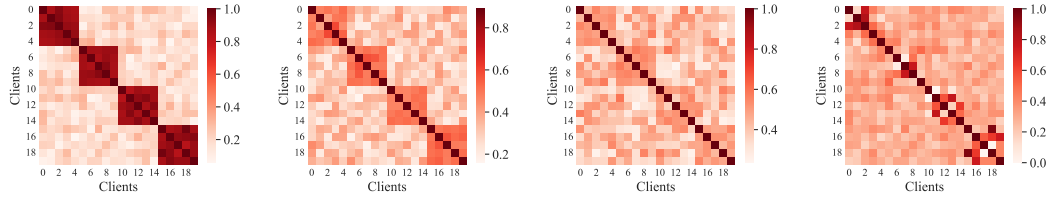
where $\mathcal{L}(\Psi) := \sum_{k=1}^K p_k \mathcal{L}_k(\Psi_k)$ is the global objective with client sampling probability p_k (w.l.o.g., $p_k = 1/K$). $\mathcal{L}^* := \sum_k p_k \mathcal{L}_k^*$ is the weighted optimal value.

The proof is in Appendix C.3. In Theorem 3.3, the first term decays linearly; the second is the classical SGD variance term; the third is the personalization error and vanishes as $\rho \rightarrow 0$. Thus FedAux can linearly descend to a neighborhood of the global optimum.

4 Experiments

4.1 Experimental Setup

Datasets and Experimental Settings Following previous works [40, 3, 42], we construct distributed subgraphs from benchmark datasets by partitioning each original graph into multiple subgraphs corresponding to individual clients. Specifically, we perform experiments on six widely used datasets, including four citation networks (Cora, CiteSeer, Pubmed [25], and ogbn-arxiv [10]) and two product co-purchase networks (Amazon-Computer and Amazon-Photo [19, 26]). We employ METIS [12] as our default graph partitioning algorithm, which allows explicit specification of the desired number of subgraphs without overlapping. Based on these datasets, we follow the standard experimental setup used in personalized subgraph federated learning literature [3, 42]. Specifically,



(a) Embedding Similarity (b) APV-based Similarity (c) Weight Similarity (d) Functional Similarity

Figure 3: Client similarity based on different measures. Darker colors indicate higher similarity.

for dataset splitting, we randomly sample 20%/40%/40% of nodes from each subgraph for training, validation, and testing, respectively. The only exception is the ogbn-arxiv dataset, due to its significantly larger size. For this dataset, we randomly select 5% of the nodes for training, half of the remaining nodes for validation, and the rest for testing. Dataset statistics and implementation details can be found in Appendix D.

Baselines FedAux is compared against several representative federated learning (FL) methods, including general FL baselines: FedAvg [20], FedProx [14], and FedPer [2]; as well as specialized graph-based FL models¹: GCFL [35], FedGNN [34], FedGTA [16], FedSage+[40], and FED-PUB[3]. Additionally, we consider a local variant of our model (*Local*), where FedAux is trained independently at each client without parameter sharing.

4.2 Main Results

As summarized in Table 1, FedAux is the only algorithm that wins every dataset–client-count combination, indicating that the APV-driven personalization generalizes from small citation graphs (Cora, CiteSeer) to large-scale, high-dimensional graphs (ogbn-arxiv). Specifically, relative to the strongest competitor in each column, FedAux achieves accuracy improvements ranging from 0.2% to 2.4%. The margin over the canonical FedAvg is more pronounced, with an average gain of 4.5%, underscoring that the proposed auxiliary projection mechanism confers benefits well beyond classical weighted averaging. Further, to quantify statistical heterogeneity (i.e., degree of non-IIDness), we adopt $\xi = \text{JSD} + \text{MMD}$ where the Jensen–Shannon Divergence (JSD) captures label-distribution skew and the Maximum Mean Discrepancy (MMD) captures disparities in subgraph structure (formal definition in Appendix D.3). Higher values of ξ indicate stronger non-IIDness. As reported in Table 5 of Appendix D.3, ξ rises monotonically for every dataset as the federation enlarges from 5 to 20 clients, confirming that our partition protocol indeed induces progressively harsher heterogeneity. This trend is mirrored in the performance of all methods, whose accuracies decline with larger client counts. Nevertheless, the drop for FedAux is consistently the smallest: on Cora, accuracy falls by only 2.9%, while FedAvg and FedProx lose nearly 7%. To highlight the contrast, we single out the most IID dataset Pubmed and the most non-IID dataset Amazon-Photo in Table 2. The results under these settings show that FedAux remains the top performer in both extremes, indicating that our model is merely tuned for gentle partitions but retains its edge under pronounced non-IID conditions.

Table 2: Degree of non-IIDness. Pubmed exhibits the lowest non-IIDness, and Amazon-Photo has the highest.

Non-IIDness	Pubmed		
	5 Clients	10 Clients	20 Clients
ξ	0.1316	0.1500	0.1725
Non-IIDness	Amazon-Photo		
	5 Clients	10 Clients	20 Clients
ξ	0.3398	0.3668	0.4307

4.3 Model Analysis

Effectiveness of APV-based Client Similarity Estimation To intuitively show that the auxiliary projection vector APV can accurately capture the latent similarity among clients under non-IIDness, we construct a synthetic graph (See Appendix D.4) that jointly embodies the two types of heterogeneity:

¹We exclude weakly privacy-protected baselines such as FedGCN [37], GraphFL [32], and FedStar [30] as these methods leak node embeddings, connectivities, or local gradients.

Table 3: Attack AUC of MIA.

Dataset	FedGNN	FedGTA	FedAux
Cora	0.56	0.54	0.51
Pubmed	0.58	0.55	0.52
ogbn-arxiv	0.55	0.53	0.49

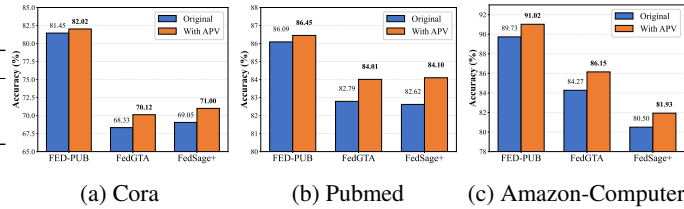


Figure 4: Transferability of APV.

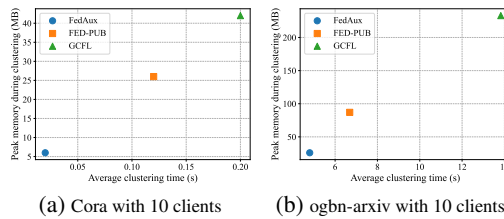


Figure 5: Server-side clustering cost.

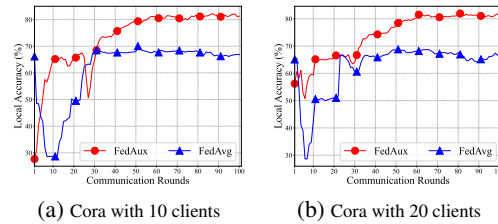


Figure 6: Averaged local accuracy concerning communication rounds.

label heterogeneity and subgraph (structural and feature) heterogeneity. Results in Fig. 3 demonstrate that the APV-based client similarity can best recover the ground-truth client relationships. Each APV converges to the principal axis of its client’s embedding distribution, yielding highly aligned directions within the same group and near-orthogonal directions across different groups (Theorem 3.1). Thus APV can serve as a privacy-preserving yet information-rich descriptor in subgraph FL.

Membership Inference Attacks (MIA) We add a comprehensive empirical privacy evaluation via MIAs to rigorously test whether our APV-based framework leaks sensitive information. We assume an honest-but-curious server with full access to the entire history of a client’s APV $\mathbf{a}_k$ and its GNN parameters θ_k . The goal of the server is to decide whether a probe embedding h_{prob} originated from client G_k , using only $(\mathbf{a}_k, \theta_k)$, while raw data, node embeddings, and gradients are never shared. Following the standard MIA methodology [21, 27], the server trains an attack classifier $g_k(\mathbf{a}_k, \theta_k, h_{\text{prob}})$ implemented as a two-layer MLP with hidden size 64, to output the probability that h_{prob} belongs to the training set of G_k . The attack is trained by maximum likelihood on a held-out mixture of member against non-member probes. In Table 3, we compare FedAux against representative subgraph FL baselines on three datasets using identical client partitions to those in our main experiments, and report the Attack AUC (lower values indicate better privacy) averaged over five random seeds. The results show that MIAs against FedAux achieve AUCs in the range [0.49, 0.52], which is indistinguishable from random guessing. This demonstrates that APVs do not leak sensitive membership information and actually provide stronger privacy than baselines.

Transferability of APV To examine whether APV can generalize beyond FedAux and serve as a plug-and-play personalization module, we integrate it into representative subgraph FL baselines. Specifically, we replace FED-PUB’s functional embedding similarity with APV, and jointly train APV with FedGTA and FedSage+. All variants are trained under the same 10-client non-IID split on three datasets. Fig. 4 shows that APV consistently improves each baseline across datasets. For example, FedGTA with APV achieves 1%-2% higher accuracy than its original counterpart. These improvements confirm that APV functions as a generic, lightweight, and privacy-preserving similarity proxy that can be seamlessly integrated with diverse subgraph FL frameworks.

Clustering Efficiency on Server To perform personalized FL, FED-PUB uses soft clustering by running a proxy graph through each client model and comparing embeddings. GCFL applies hard clustering, grouping clients with a Stoer-Wagner cut on cross-round gradients. Our APV-based method only computes similarity between the uploaded vectors, requiring no extra forward passes, gradient logging or cut computation, so server overhead is minimal. As Fig. 5 illustrates, the APV method attains the fastest clustering time and the lowest peak memory, making it more efficient at scale.

Convergence Rates Fig. 6 compares the convergence behavior of FedAux and FedAvg. Notably, FedAux converges by the 60th communication round in both the 10-client and 20-client settings. Since the latter involves a higher level of data heterogeneity, it indicates that FedAux maintains a consistent convergence speed even as the degree of non-IIDness increases.

In Appendix E, we conduct an ablation study to investigate the impact of our proposed continuous aggregation scheme, which encodes local node relationships in the 1D space induced by the APV, as well as the effect of the server-side APV-based personalized federated aggregation. We also conduct experiments to analyze the sensitivity to hyperparameters.

5 Related Work

For general **Federated Learning**, FedAvg [20] first demonstrated that deep models can be trained on decentralized data with iterative model averaging. Subsequent work revealed that statistical and systems heterogeneity slow or destabilize FedAvg’s convergence, which has been formally addressed by proximal correction [14], variance reduction [11], and data–distribution smoothing through a small globally shared subset of samples [43]. Optimization refinements such as normalized aggregation [33], adaptive server updates [24] and dynamic regularization [1] further tighten convergence guarantees under extreme non-IID settings. Beyond a single global model, personalization frameworks, e.g., Ditto [15], which jointly optimizes a shared model and client-specific objectives, explicitly trade off fairness, robustness, and local adaptation. These advances establish the algorithmic and theoretical foundations on which federated graph learning is built.

For **Graph Federated Learning** [36, 17, 39, 38], at the node level, FedGCN [37] illustrates that one-shot encrypted exchange can suffice to federate GCNs while maintaining accuracy and privacy; GCFL [35] clusters clients by gradient dynamics to mitigate structural and feature shift across graphs. To combat cross-domain heterogeneity, FedStar [30] extracts a domain-invariant topology that generalizes across diverse graphs, and FedGraph [9] augments local data by requesting node information from other clients. When each participant owns only a fragment of a larger network, subgraph federated learning methods such as FedSage/FedSage+ [40] generate virtual neighbours to repair missing cross-subgraph edges. FED-PUB [3] proposes to generate functional embeddings to evaluate the similarity between clients for personalized aggregation. These models collectively highlight an open challenge for personalized graph FL, i.e., accurate client similarity measures, which our proposed FedAux addresses through end-to-end learning of auxiliary projection vectors.

Several new methods further extend personalization but come with non-trivial trade-offs. FedSSP [31] transmits spectral components that are invariant across domains but can leak sensitive local spectral information. FedEgo [41] shares ego-network embeddings, which directly expose structural patterns. PFGNAS [8] leverages prompt-based neural architecture search without server-side similarity estimation, but its reliance on LLM-based personalization introduces prohibitive cost. FedGrAINS [5] personalizes within each client by learning with a GFlowNet, thereby avoiding server-side mixing, but at the expense of training an additional model with high compute and memory overhead. While the aforementioned methods contribute interesting ideas, they either compromise privacy guarantees (e.g., FedSSP, FedEgo) or impose heavy computational costs (e.g., PFGNAS, FedGrAINS), limiting their practicality. In contrast, our method learns Auxiliary Projection Vectors (APV) in an end-to-end manner: lightweight, privacy-preserving client signatures that avoid raw data or embedding leakage, yet enable effective similarity-aware aggregation for personalized subgraph FL.

6 Conclusion

We present FedAux, a personalized subgraph federated learning framework that augments each local GNN with learnable auxiliary projection vectors (APVs). Specifically, besides the global GNN parameters, the server initializes and distributes APVs to each client, enabling effective and privacy-preserving characterization of local subgraph structures. By continuously projecting node embeddings onto a 1D space induced by these APVs, local models adaptively refine the APVs to optimally capture node relationships within each client’s subgraph. After each communication round, the server leverages similarity between client APVs to perform personalized model aggregation. Extensive experiments across multiple datasets with varying numbers of clients validate the effectiveness of APVs as informative descriptors for personalized subgraph FL.

Acknowledgements

The research is supported, in part, by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (RG101/24).

References

- [1] Acar, D. A. E., Zhao, Y., Matas, R., Mattina, M., Whatmough, P., and Saligrama, V. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.
- [2] Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.
- [3] Baek, J., Jeong, W., Jin, J., Yoon, J., and Hwang, S. J. Personalized subgraph federated learning. In *International conference on machine learning*, pp. 1396–1415. PMLR, 2023.
- [4] Bellman, R. Dynamic programming. *science*, 153(3731):34–37, 1966.
- [5] Ceyani, E., Xie, H., Buyukates, B., Yang, C., and Avestimehr, S. Fedgrains: Personalized subgraph federated learning with adaptive neighbor sampling. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*, pp. 598–607. SIAM, 2025.
- [6] Csiszár, I. and Körner, J. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.
- [7] Delfour, M. C. and Zolésio, J.-P. *Shapes and geometries: metrics, analysis, differential calculus, and optimization*. SIAM, 2011.
- [8] Fang, H., Gao, Y., Zhang, P., Yao, J., Chen, H., and Wang, H. Large language models enhanced personalized graph neural architecture search in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16514–16522, 2025.
- [9] He, C., Balasubramanian, K., Ceyani, E., Yang, C., Xie, H., Sun, L., He, L., Yang, L., Yu, P. S., Rong, Y., Zhao, P., Huang, J., Annamaram, M., and Avestimehr, S. Fedgraphnn: A federated learning system and benchmark for graph neural networks, 2021.
- [10] Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.
- [11] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.
- [12] Karypis, G. Metis: Unstructured graph partitioning and sparse matrix ordering system. *Technical report*, 1997.
- [13] Krasulina, T. The method of stochastic approximation for the determination of the least eigenvalue of a symmetrical matrix. *USSR Computational Mathematics and Mathematical Physics*, 9(6):189–195, 1969.
- [14] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.
- [15] Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pp. 6357–6368. PMLR, 2021.
- [16] Li, X., Wu, Z., Zhang, W., Zhu, Y., Li, R.-H., and Wang, G. Fedgta: Topology-aware averaging for federated graph learning. *Proc. VLDB Endow.*, pp. 41–50, 2023.

- [17] Li, Z., Wang, X., Chen, H.-Y., Shen, H. W., and Chao, W.-L. H. Fedne: Surrogate-assisted federated neighbor embedding for dimensionality reduction. *Advances in Neural Information Processing Systems*, 37:133948–133974, 2024.
- [18] Liu, M., Wang, Z., and Ji, S. Non-local graph neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):10270–10276, 2021.
- [19] McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 43–52, 2015.
- [20] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.
- [21] Nasr, M., Shokri, R., and Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753. IEEE, 2019.
- [22] Oja, E. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15:267–273, 1982.
- [23] Polyak, B. T. Gradient methods for minimizing functionals. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, 3(4):643–653, 1963.
- [24] Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.
- [25] Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.
- [26] Shchur, O., Mumme, M., Bojchevski, A., and Günnemann, S. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.
- [27] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.
- [28] Suzumura, T., Zhou, Y., Baracaldo, N., Ye, G., Houck, K., Kawahara, R., Anwar, A., Stavarache, L. L., Watanabe, Y., Loyola, P., et al. Towards federated graph learning for collaborative financial crimes detection. *arXiv preprint arXiv:1909.12946*, 2019.
- [29] Tan, A. Z., Yu, H., Cui, L., and Yang, Q. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022.
- [30] Tan, Y., Liu, Y., Long, G., Jiang, J., Lu, Q., and Zhang, C. Federated learning on non-iid graphs via structural knowledge sharing. In *Proceedings of the AAAI conference on artificial intelligence*, 2023.
- [31] Tan, Z., Wan, G., Huang, W., and Ye, M. Fedssp: Federated graph learning with spectral knowledge and personalized preference. *Advances in Neural Information Processing Systems*, 37:34561–34581, 2024.
- [32] Wang, B., Li, A., Pang, M., Li, H., and Chen, Y. Graphfl: A federated learning framework for semi-supervised node classification on graphs. In *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 498–507. IEEE, 2022.
- [33] Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.
- [34] Wu, C., Wu, F., Cao, Y., Huang, Y., and Xie, X. Fedgmn: Federated graph neural network for privacy-preserving recommendation. *arXiv preprint arXiv:2102.04925*, 2021.

- [35] Xie, H., Ma, J., Xiong, L., and Yang, C. Federated graph classification over non-iid graphs. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.
- [36] Xie, H., Xiong, L., and Yang, C. Federated node classification over graphs with latent link-type heterogeneity. In *Proceedings of the ACM Web Conference 2023*, pp. 556–566, 2023.
- [37] Yao, Y., Jin, W., Ravi, S., and Joe-Wong, C. Fedgcn: convergence-communication tradeoffs in federated training of graph convolutional networks. *Advances in neural information processing systems*, 2023.
- [38] Ye, R., Ni, Z., Wu, F., Chen, S., and Wang, Y. Personalized federated learning with inferred collaboration graphs. In *International Conference on Machine Learning*, pp. 39801–39817. PMLR, 2023.
- [39] Yue, L., Liu, Q., Gao, W., Liu, Y., Zhang, K., Du, Y., Wang, L., and Yao, F. Federated self-explaining gnns with anti-shortcut augmentations. In *Forty-first International Conference on Machine Learning*, 2024.
- [40] Zhang, K., Yang, C., Li, X., Sun, L., and Yiu, S. M. Subgraph federated learning with missing neighbor generation. *Advances in Neural Information Processing Systems*, 34:6671–6682, 2021.
- [41] Zhang, T., Mai, C., Chang, Y., Chen, C., Shu, L., and Zheng, Z. Fedego: privacy-preserving personalized federated graph learning with ego-graphs. *ACM Transactions on Knowledge Discovery from Data*, 18(2):1–27, 2023.
- [42] Zhang, Z., Hu, Q., Yu, Y., Gao, W., and Liu, Q. Fedgt: federated node classification with scalable graph transformer. *arXiv preprint arXiv:2401.15203*, 2024.
- [43] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.
- [44] Zhuo, W. and Tan, G. Efficient graph similarity computation with alignment regularization. *Advances in Neural Information Processing Systems*, 35:30181–30193, 2022.
- [45] Zhuo, W., Liu, Z., Hooi, B., He, B., Tan, G., Fathony, R., and Chen, J. Partitioning message passing for graph fraud detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=tEgrUrUuWA>.
- [46] Zhuo, W., Yu, H., Tan, G., and Li, X. Commute graph neural networks. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=29Leye9511>.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes] .

Justification: In the abstract and Introduction, we have highlighted the contributions of our work, such as lines 54 - 70.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes] .

Justification: See Appendix F.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes] .

Justification: All theorems are in Section 3.3. All assumptions and proofs are shown in Appendix C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes] .

Justification: We provide the implementation details in Appendix D. Besides, we provide the source code of our work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#) .

Justification: We have added the link to the source code of our work. Besides, we provide a detailed README file to reproduce our experimental results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#) .

Justification: See Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#) .

Justification: See Table 1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#) .

Justification: See Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#) .

Justification: Our research conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[NA\]](#) .

Justification: There are no negative societal impacts of our work.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA] .

Justification: No risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes] .

Justification: All datasets used in this paper are public, and we have properly cited the original paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA] .

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA] .

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA] .

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A More Discussion of Related Work

In subgraph federated learning, each client holds a local subgraph G_k of a global graph G . These subgraphs can vary substantially in their feature distributions, structural/topological properties, and label distributions. Thus, simply applying FedAvg may fail to converge properly or yield a suboptimal global model, because clients often learn different parameters tailored to their local subgraphs, and blindly averaging those parameters disregards the non-IID nature of the data. To address this, recent studies move beyond basic FedAvg by introducing personalized or locality-aware aggregation schemes that better handle heterogeneous subgraph data.

GCFL [35] compares each client's parameter updates before and after communication. If a client's update deviates substantially from the majority, it is deemed too different and is excluded from the dominant aggregation cluster. However, GCFL relies on hard clustering, which cannot capture finer-grained similarities across clients. Moreover, it depends on manually tuned hyperparameters to control how different a client must be before exclusion, leaving it unclear how large a deviation is tolerable without enough knowledge about the raw data. FED-PUB [3], on the other hand, constructs a proxy graph at the server and evaluates model outputs from each client on the proxy graph to measure their similarity. Clients with similar outputs on the proxy graph are deemed more alike and thus aggregated more closely. Yet building a suitable proxy graph as the middleware to measure the similarity between clients is nontrivial, because it requires simulating the real graph data on the server.

In this work, we propose an orthogonal approach to measure and aggregate models on the server. Specifically, beyond distributing a global GNN, the server also provides a global Auxiliary Projection Vector (APV) to each client at the start of every communication round, as shown on the left of Fig. 1. During local training, the APV is jointly optimized with the client's GNN parameters to form a one-dimensional space onto which node embeddings are projected. This procedure tailors the APV, which adjusts the shape of this space, to the unique distribution of each subgraph by preserving its distinctive patterns. Once training concludes, clients upload their updated GNN parameters and local APV to the server, which then compares these learned APVs to gauge similarity and detect finer-grained, continuous heterogeneity across clients. By discarding the hard clustering thresholds used in GCFL and removing the need for a proxy dataset used in FED-PUB, our method offers a more flexible and efficient strategy to identify and aggregate similar subgraphs.

Another advantage of our method lies in its privacy-preserving design. We argue that sharing detailed node embeddings, as in FedGCN [37], can leak private subgraph information, as adversaries on the server could compute pairwise similarities among embeddings to reconstruct local connectivity. Moreover, due to high dimensionality and lack of explicit structural encoding, directly comparing parameter matrices to measure raw data similarity is both unreliable and computationally unstable. In contrast, our proposed APV is a compact parameter vector that preserves essential node relationships without exposing the actual node embeddings, offering stronger privacy guarantees and structure-awareness. Additionally, its low-dimensional nature makes similarity computation more efficient and robust.

B Pseudo Code of FedAux

The overall training algorithm is shown in Algorithm 1.

C Proofs

C.1 Proof of Theorem 3.1

The proof relies on two mild structural assumptions, both typical in optimization analyses of deep models.

Assumption C.1 (Centred embeddings). $\bar{h} := \frac{1}{N} \sum_i h_i = \mathbf{0}$, where N is the number of nodes in the current client.

Assumption C.2 (Linear classifier Jacobian). Let $g_i := \nabla_{z_i} \mathcal{L}(\theta, \mathbf{a})$. Assuming that the mapping $z_i \mapsto g_i$ is linear: $g_i = \mathbf{W} z_i$ for some matrix $\mathbf{W} \in \mathbb{R}^{d' \times d'}$.

Algorithm 1: FedAux: Subgraph FL with Differentiable Auxiliary Projections

Input: number of clients K ; global communication rounds T ; local steps per round Q ; learning rates η ; similarity temperature α

Init: server initializes global GNN weights $\theta^{(0)}$ and APV $\mathbf{a}^{(0)} \sim \mathcal{N}(0, I_{d'})$; each client G_k sets $\theta_k^{(0)} \leftarrow \theta^{(0)}, \mathbf{a}_k^{(0)} \leftarrow \mathbf{a}^{(0)}$

```
1 for  $t \leftarrow 1$  to  $T$  do
  // Server  $\rightarrow$  Clients
2 broadcast  $\{(\theta_k^{(t-1)}, \mathbf{a}_k^{(t-1)})\}_{k=1}^K$  to each client  $\{G_k\}_{k=1}^K$ ;
  // Local training on each client  $G_k$  (runs in parallel)
3 for each client  $G_k \in \{G_1, \dots, G_K\}$  do in parallel
4    $(\theta_k, \mathbf{a}_k) \leftarrow (\theta_k^{(t-1)}, \mathbf{a}_k^{(t-1)})$ ;
5   for  $q \leftarrow 1$  to  $Q$  do
6     run local GNN forward pass with  $h_{k,i}^{(t-1)} = f_{\theta_k^{(t-1)}}(v_i)$  to learn node embeddings as
       Eq. (1);
7     compute similarity scores  $s_k^{(t-1)}$  with  $s_{k,i}^{(t-1)} = \langle \hat{h}_{k,i}^{(t-1)}, \mathbf{a}_k^{(t-1)} \rangle$ ;
8     compute kernel weights  $\kappa(s_{k,i}^{(t-1)}, s_{k,j}^{(t-1)})$ ;
9     compute kernel-based aggregated embeddings  $z_{k,i}^{(t-1)}$  with Eq. (6);
10    concatenate  $r_{k,i}^{(t-1)} = [h_{k,i}^{(t-1)} \| z_{k,i}^{(t-1)}]$ ;
11    forward through the classifier; compute loss  $\mathcal{L}_{CE}$ ;
12    update  $(\theta_k^{(t-1)}, \mathbf{a}_k^{(t-1)})$  w.r.t.  $\mathcal{L}_{CE}$  with learning rate  $\eta$ ;
13  end
14  upload  $(\theta_k^{(t-1)}, \mathbf{a}_k^{(t-1)})$  to server;
15 end
  // Server-side personalised aggregation
16 compute  $w_{k,l}^{(t-1)}$  with Eq. (7);
17 for each client  $G_k$  do
18    $\theta_k^{(t)} \leftarrow \sum_{l=1}^K w_{k,l}^{(t-1)} \theta_l^{(t-1)}$ ;
19    $\mathbf{a}_k^{(t)} \leftarrow \sum_{l=1}^K w_{k,l}^{(t-1)} \mathbf{a}_l^{(t-1)}$ ;
20 end
21 send  $(\theta_k^{(t)}, \mathbf{a}_k^{(t)})$  back to client  $G_k$ ;
22 end
```

Output: personalized models $\{\theta_k^{(T)}, \mathbf{a}_k^{(T)}\}_{k=1}^K$ for K clients

Theorem C.2 holds exactly for a linear-softmax classifier and is a first-order approximation for MLPs. Then we introduce the auxiliary lemmata.

Lemma C.3 (Gradient of similarity score). *For node v_i , $\frac{\partial s_i}{\partial \mathbf{a}} = h_i - s_i \mathbf{a}$ holds.*

Proof. $s_i = \mathbf{a}^\top h_i$, with $\|\mathbf{a}\| = 1$. Hence $\frac{\partial s_i}{\partial \mathbf{a}} = h_i$. Because we will always re-project $\mathbf{a}$ on the unit sphere after every update, the tangential component $h_i - s_i \mathbf{a}$ is the effective gradient, and the radial component vanishes. $\square$

Lemma C.4 (Gradient of kernel entry). $\frac{\partial \mathcal{K}_{ij}}{\partial \mathbf{a}} = -\frac{2}{\sigma^2} (s_i - s_j) \mathcal{K}_{ij} [(h_i - h_j) - (s_i - s_j) \mathbf{a}]$.

Proof. Apply the chain rule to $\mathcal{K}_{ij} = \kappa(s_i, s_j) = \exp\left(-\frac{1}{\sigma^2} (s_i - s_j)^2\right)$ and invoke Theorem C.3 for $\partial(s_i - s_j)/\partial \mathbf{a}$. $\square$

Lemma C.5 (Gradient of the kernel-smoothed embedding). *Define the normalized kernel weights:*

$$\beta_{ij} := \frac{\mathcal{K}_{ij}}{M_i}, \quad \sum_j \beta_{ij} = 1. \quad (13)$$

Then we have

$$\frac{\partial z_i}{\partial \mathbf{a}} = -\frac{2}{\sigma^2} \sum_{j=1}^N \beta_{ij} (s_i - s_j) (h_j - z_i) \otimes [(h_i - h_j) - (s_i - s_j) \mathbf{a}], \quad (14)$$

where $\otimes$ denotes the outer product.

Proof. Let $D_i = \sum_{j=1}^N \mathcal{K}_{ij} h_j$, based on Eq. (6) we have $z_i = \frac{D_i}{M_i}$. Using the quotient rule, we have:

$$\frac{\partial z_i}{\partial \mathbf{a}} = \frac{1}{M_i} \frac{\partial D_i}{\partial \mathbf{a}} - \frac{D_i}{M_i^2} \frac{\partial M_i}{\partial \mathbf{a}}. \quad (15)$$

Then we compute the two gradients in Eq. (15). The gradient of the D_i w.r.t. $\mathbf{a}$ is $\frac{\partial D_i}{\partial \mathbf{a}} = \sum_{j=1}^N \frac{\partial \mathcal{K}_{ij}}{\partial \mathbf{a}} h_j^\top$, where $\frac{\partial \mathcal{K}_{ij}}{\partial \mathbf{a}}$ has been given by Theorem C.4. Thus we have:

$$\frac{\partial D_i}{\partial \mathbf{a}} = -\frac{2}{\sigma^2} \sum_{j=1}^N (s_i - s_j) \mathcal{K}_{ij} h_j [(h_i - h_j) - (s_i - s_j) \mathbf{a}]^\top. \quad (16)$$

The Gradient of M_i can be represented as:

$$\frac{\partial M_i}{\partial \mathbf{a}} = \sum_{j=1}^N \frac{\partial \mathcal{K}_{ij}}{\partial \mathbf{a}} = -\frac{2}{\sigma^2} \sum_{j=1}^N (s_i - s_j) \mathcal{K}_{ij} [(h_i - h_j) - (s_i - s_j) \mathbf{a}]. \quad (17)$$

We can substitute Eq. (16) and Eq. (17) into the quotient rule Eq. (15):

$$\begin{aligned} \frac{\partial z_i}{\partial \mathbf{a}} = & -\frac{2}{\sigma^2} \frac{1}{M_i} \sum_{j=1}^N (s_i - s_j) \mathcal{K}_{ij} h_j [(h_i - h_j) - (s_i - s_j) \mathbf{a}]^\top \\ & + \frac{2}{\sigma^2} \frac{1}{M_i^2} \left(\sum_{j=1}^N \mathcal{K}_{ij} h_j \right) \left(\sum_{\ell=1}^N (s_i - s_\ell) \mathcal{K}_{i\ell} [(h_i - h_\ell) - (s_i - s_\ell) \mathbf{a}]^\top \right). \end{aligned} \quad (18)$$

Given Eq. (13), Eq. (18) can be rewritten as:

$$\begin{aligned} \frac{\partial z_i}{\partial \mathbf{a}} = & -\frac{2}{\sigma^2} \sum_{j=1}^N \beta_{ij} (s_i - s_j) h_j [(h_i - h_j) - (s_i - s_j) \mathbf{a}]^\top \\ & + \frac{2}{\sigma^2} \underbrace{\left(\sum_{j=1}^N \beta_{ij} h_j \right)}_{z_i} \left(\sum_{\ell=1}^N \beta_{i\ell} (s_i - s_\ell) [(h_i - h_\ell) - (s_i - s_\ell) \mathbf{a}]^\top \right). \end{aligned} \quad (19)$$

By re-indexing $\ell \rightarrow j$, Eq. (19) can be represented as:

$$\frac{\partial z_i}{\partial \mathbf{a}} = -\frac{2}{\sigma^2} \sum_{j=1}^N \beta_{ij} (s_i - s_j) (h_j - z_i) [(h_i - h_j) - (s_i - s_j) \mathbf{a}]^\top, \quad (20)$$

which is exactly the Eq. (14), completing the derivation. $\square$

C.1.1 Proof of Theorem 3.1

Proof. Using Theorem C.2 and Theorem C.5, we can perform chain-rule expansion of $\nabla_{\mathbf{a}} \mathcal{L}$ as:

$$\nabla_{\mathbf{a}} \mathcal{L} = \frac{1}{N} \sum_{i=1}^N \left(\frac{\partial z_i}{\partial \mathbf{a}} \right)^\top g_i = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \left(\frac{\partial z_i}{\partial \mathbf{a}} \right)^\top \mathbf{W} z_i. \quad (21)$$

Substituting Eq. (14) into Eq. (21) gives:

$$\nabla_{\mathbf{a}} \mathcal{L} = -\frac{2}{N\sigma^2} \sum_{i=1}^N \sum_{j=1}^N \beta_{ij} (s_i - s_j) [(h_i - h_j) - (s_i - s_j) \mathbf{a}] (h_j - z_i)^\top \mathbf{W} z_i. \quad (22)$$

For small σ , β_{ij} is sharply peaked at $j = i$. Given $\varepsilon_{ij} := s_i - s_j$, since $\beta_{ij} \leq e^{-\varepsilon_{ij}^2/\sigma^2}$, all terms with $j \neq i$ are exponentially suppressed, and the dominant contribution arises from the linearization around $\varepsilon_{ij} = 0$. Then we conduct Taylor expansion to first order in ε_{ij} :

$$\nabla_a \mathcal{L} = -\frac{2}{N\sigma^2} \sum_{i=1}^N \sum_{j=1}^N [\beta_{ij} \varepsilon_{ij} h_i (h_i^\top \mathbf{W} h_i)] + \mathcal{O}(\sigma^0). \quad (23)$$

In Big-O notation, the symbol $\mathcal{O}(\sigma^\mu)$ as $\sigma \rightarrow 0^+$ means that there exists a constant $c > 0$ and a neighbourhood $(0, \sigma_0]$ such that $|\mathcal{O}(\sigma^\mu)| \leq c\sigma^\mu$. Using Theorem C.1 and symmetry of the inner summation one obtains the compact matrix form of the above formula:

$$\nabla_a \mathcal{L} = -\frac{2}{\sigma^2} \left(\frac{1}{N} \sum_{i=1}^N h_i h_i^\top \right) \mathbf{a} + \mathcal{R}(\sigma) = -\frac{2}{\sigma^2} \mathbf{C} \mathbf{a} + \mathcal{R}(\sigma), \quad (24)$$

with $\|\mathcal{R}(\sigma)\| = \mathcal{O}(\sigma^0)$. This proves Eq. (9).

Since $\mathbf{a}$ is re-normalised after every update, the effective tangential gradient [7] is $\nabla_a \mathcal{L} - (\mathbf{a}^\top \nabla_a \mathcal{L}) \mathbf{a}$. Note that $\mathbf{a}^\top \mathbf{C} \mathbf{a}$ is scalar, so subtracting the radial part yields the tangential gradient $-\mathbf{C} \mathbf{a} + (\mathbf{a}^\top \mathbf{C} \mathbf{a}) \mathbf{a}$. A projected gradient descent step with learning rate η therefore becomes:

$$\mathbf{a} \leftarrow \Pi_{\mathbb{S}^{d-1}} (\mathbf{a} - \eta [\mathbf{C} \mathbf{a} - (\mathbf{a}^\top \mathbf{C} \mathbf{a}) \mathbf{a}]) = \Pi_{\mathbb{S}^{d-1}} ((\mathbf{I} - \eta \mathbf{C}) \mathbf{a}), \quad (25)$$

which is exactly the discrete-time Oja [22] update Eq. (10).

Let $\lambda_{\max}$ be the largest eigenvalue of $\mathbf{C}$, with unit-norm eigenvector $u_{\max}$. Standard theory of Oja's algorithm [13, 22] states:

- Every eigenvector of $\mathbf{C}$ is a fixed point of Eq. (10).
- All eigenvectors other than $\pm u_{\max}$ are unstable, and $\pm u_{\max}$ are globally asymptotically stable provided $0 < \eta < 2/\lambda_{\max}$.

Hence gradient descent drives $\mathbf{a}$ toward $\pm u_{\max}$. Because the kernel and the classifier do not depend on the sign of $\mathbf{a}$, both directions are equivalent, and choosing the positive-projection suffices. $\square$

C.1.2 Interpretation of Theorem 3.1

Rayleigh-quotient maximization Oja learning rule is a stochastic gradient ascent on the Rayleigh quotient $\mathcal{R}(\mathbf{a}) = \mathbf{a}^\top \mathbf{C} \mathbf{a}$ over the unit sphere. Theorem 3.1 therefore formalizes the intuition that the APV aligns with the **direction of maximum embedding variance**.

Role of the bandwidth σ The leading term Eq. (10) is multiplied by $1/\sigma^2$. A smaller bandwidth increases the gradient magnitude, accelerating alignment but reducing smoothness; conversely, a larger σ slows convergence while preserving differentiability.

Compatibility with the global optimization Once the APV converges to $u_{\max}$, the kernel weights $\mathcal{K}_{ij}$ depend only on $\langle h_i, u_{\max} \rangle - \langle h_j, u_{\max} \rangle$, which maximally separates nodes along the most informative one-dimensional projection. This precisely captures the fidelity property we desire.

C.2 Proof of Theorem 3.2

The proof is based on three lemmata.

Lemma C.6 (Weight concentration). *For each row i of the kernel matrix $\mathcal{K}$, we have:*

$$\lim_{\sigma \rightarrow 0^+} \mathcal{K}_{ij} = \delta_{ij}, \quad (26)$$

where δ_{ij} is the Kronecker delta.

Proof. Since all scores are distinct, setting $\Delta_i := \min_{j \neq i} |s_i - s_j| > 0$. For $j \neq i$ we have:

$$\frac{\mathcal{K}_{ij}}{\mathcal{K}_{ii}} = \exp \left(-\frac{((s_i - s_j)^2 - 0)}{\sigma^2} \right) \leq \exp \left(-\frac{\Delta_i^2}{\sigma^2} \right). \quad (27)$$

Hence for all $j \neq i$, $\mathcal{K}_{ij} \leq e^{-\Delta_i^2/\sigma^2} \rightarrow 0$ as $\sigma \rightarrow 0^+$. Since each row of $\mathcal{K}$ is a probability distribution, the diagonal entry must satisfy $\mathcal{K}_{ii} = 1 - \sum_{j \neq i} \mathcal{K}_{ij} \rightarrow 1$. $\square$

Lemma C.7 (Pointwise convergence of smoothed embeddings).

$$\lim_{\sigma \rightarrow 0^+} z_i = h_i, \quad \forall i = 1, \dots, N \quad (28)$$

Proof. We can rewrite z_i as:

$$z_i = \sum_{j=1}^N \alpha_{ij} h_j = \mathcal{K}_{ii} h_i + \sum_{j \neq i} \mathcal{K}_{ij} h_j. \quad (29)$$

By Theorem C.6 the non-diagonal weights vanish and $\mathcal{K}_{ii} \rightarrow 1$. Therefore $z_i \rightarrow h_i$. $\square$

Lemma C.8 (Matrix convergence in Frobenius norm).

$$\lim_{\sigma \rightarrow 0^+} \left\| \tilde{\mathbf{Z}} - \tilde{\mathbf{H}} \right\|_F = 0. \quad (30)$$

Proof. Note that both matrices $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{H}}$ have the same ordering π of rows. From Theorem C.7 each corresponding row converges as $\|z_{\pi(t)} - h_{\pi(t)}\|_2 \rightarrow 0$ for every t . Since N is finite, the Frobenius norm also converges to 0. $\square$

C.2.1 Proof of Theorem 3.2

Proof. The Conv1D operator $\text{Conv}_{\mathbf{W}}$ is linear and its Lipschitz constant with respect to the Frobenius norm is $\text{LIP}_{\mathbf{W}} = \left(\sum_{\tau=1}^B \|\mathbf{W}_{\tau}\|_2^2 \right)^{1/2}$, then for any two sequences $\mathcal{X}$ and $\mathcal{Y}$ we have:

$$\|\text{Conv}_{\mathbf{W}}(\mathcal{X}) - \text{Conv}_{\mathbf{W}}(\mathcal{Y})\|_F \leq \text{LIP}_{\mathbf{W}} \|\mathcal{X} - \mathcal{Y}\|_F. \quad (31)$$

Applying this bound with $\mathcal{X} = \tilde{\mathbf{Z}}$ and $\mathcal{Y} = \tilde{\mathbf{H}}$ yields:

$$\left\| \text{Conv}_{\mathbf{W}}(\tilde{\mathbf{Z}}) - \text{Conv}_{\mathbf{W}}(\tilde{\mathbf{H}}) \right\|_F \leq \text{LIP}_{\mathbf{W}} \|\tilde{\mathbf{Z}} - \tilde{\mathbf{H}}\|_F. \quad (32)$$

Theorem C.8 states that the right-hand side of Eq. (32) converges to 0, thus the left-hand side must converge to zero as well, establishing the claimed limit. $\square$

C.3 Proof of Theorem 3.3

C.3.1 Local Training Analysis

For a client G_k at any communication round and inner step $q \in \{1, \dots, Q\}$, let $\Psi_k^q := (\theta_k^q, \mathbf{a}_k^q)$ at any communication round. Given the assumption that the local objective $\mathcal{L}_k$ is differentiable and $\mathcal{L}$ -smooth: $\forall \Psi_k, \Psi'_k : \|\nabla \mathcal{L}_k(\Psi_k) - \nabla \mathcal{L}_k(\Psi'_k)\| \leq \mathcal{L} \|\Psi_k - \Psi'_k\|$, where the smoothness holds for cross-entropy composed with neural networks whose activations are Lipschitz, we have:

$$\mathcal{L}_k(\Psi'_k) \leq \mathcal{L}_k(\Psi_k) + \langle \nabla \mathcal{L}_k(\Psi_k), \Psi'_k - \Psi_k \rangle + \frac{\mathcal{L}}{2} \|\Psi'_k - \Psi_k\|^2, \quad \forall \Psi_k, \Psi'_k. \quad (33)$$

Take $\Psi_k = \Psi_k^{s-1}$, and $\Psi'_k = \Psi_k^q = \Psi_k^{q-1} - \eta g_k^{q-1}$:

$$\mathcal{L}_k(\Psi_k^q) \leq \mathcal{L}_k(\Psi_k^{q-1}) - \eta \langle \nabla \mathcal{L}_k(\Psi_k^{q-1}), g_k^{q-1} \rangle + \frac{\mathcal{L} \eta^2}{2} \|g_k^{q-1}\|^2. \quad (34)$$

Due to the unbiasedness assumption, i.e., $\mathbb{E}[g_k] = \nabla \mathcal{L}_k$, we have:

$$\mathbb{E}[\langle \nabla \mathcal{L}_k, g_k^s \rangle] = \langle \nabla \mathcal{L}_k, \mathbb{E}[g_k^q] \rangle = \|\nabla \mathcal{L}_k\|^2. \quad (35)$$

Also by the bounded-variance assumption, i.e., $\mathbb{E}_q [\|g_k - \nabla \mathcal{L}_k\|^2] \leq \zeta^2$, we have:

$$\mathbb{E}_q [\|g_k^q\|^2] = \|\nabla \mathcal{L}_k\|^2 + \mathbb{E} [\|g_k^q - \nabla \mathcal{L}_k\|^2] \stackrel{\text{Ass. 2}}{\leq} \|\nabla \mathcal{L}_k\|^2 + \zeta^2. \quad (36)$$

Then we can insert Eq. (35) and Eq. (36) into Eq. (34) and take expectation as follows:

$$\begin{aligned} \mathbb{E}_q [\mathcal{L}_k(\Psi_k^q)] &\leq \mathcal{L}_k(\Psi_k^{q-1}) - \eta \|\nabla \mathcal{L}_k(\Psi_k^{q-1})\|^2 + \frac{\mathcal{L}\eta^2}{2} \left(\|\nabla \mathcal{L}_k(\Psi_k^{q-1})\|^2 + \zeta^2 \right) \\ &= \mathcal{L}_k(\Psi_k^{q-1}) - \left(\eta - \frac{\mathcal{L}\eta^2}{2} \right) \|\nabla \mathcal{L}_k\|^2 + \frac{\mathcal{L}\eta^2\zeta^2}{2}. \end{aligned} \quad (37)$$

Because $\eta \leq 1/2\mathcal{L}$ we have $1 - \mathcal{L}\eta/2 \geq 1/2$, hence:

$$\mathbb{E}_q [\mathcal{L}_k(\Psi_k^q)] \leq \mathcal{L}_k(\Psi_k^{q-1}) - \frac{\eta}{2} \|\nabla \mathcal{L}_k(\Psi_k^{q-1})\|^2 + \frac{\mathcal{L}\eta^2\zeta^2}{2}. \quad (38)$$

Based on the assumption that each local objective $\mathcal{L}_k$ satisfies the μ -PL² (Polyak-Lojasiewicz) condition [23] iff

$$\|\nabla \mathcal{L}_k(\Psi_k^{q-1})\|^2 \geq 2\mu (\mathcal{L}_k(\Psi_k^{q-1}) - \mathcal{L}_k^*). \quad (39)$$

Plug Eq. (39) into Eq. (38):

$$\mathbb{E}_q [\mathcal{L}_k(\Psi_k^q) - \mathcal{L}_k^*] \leq (1 - \eta\mu) (\mathcal{L}_k(\Psi_k^{q-1}) - \mathcal{L}_k^*) + \frac{\mathcal{L}\eta^2\zeta^2}{2}. \quad (40)$$

We define the gap as $\Delta_k^{q-1} := \mathcal{L}_k(\Psi_k^{q-1}) - \mathcal{L}_k^*$. Taking the total expectation and iterating Eq. (40) Q times, we have:

$$\mathbb{E} [\Delta_k^Q] \leq (1 - \eta\mu)^Q \Delta_k^1 + \frac{\mathcal{L}\eta^2\zeta^2}{2} \sum_{j=1}^Q (1 - \eta\mu)^{j-1}. \quad (41)$$

The geometric sum is:

$$\sum_{j=1}^Q (1 - \eta\mu)^{j-1} = \frac{1 - (1 - \eta\mu)^Q}{\eta\mu} \leq \frac{1}{\eta\mu}. \quad (42)$$

Therefore, the following inequality holds:

$$\mathbb{E} [\Delta_k^Q] \leq (1 - \eta\mu)^Q \Delta_k^1 + \frac{\eta\mathcal{L}\zeta^2}{2\mu}, \quad (43)$$

which is the per-client local-training contraction.

C.3.2 Effect of Global Kernel-Based Aggregation

Define $\Psi_k^{\text{loc},(t-1)} := \Psi_k^Q$ which means the local client parameters after Q local training iterations. Let $\mathbf{f}^{(t-1)} := [f_1^{(t-1)}, \dots, f_K^{(t-1)}]^\top$, where $f_k^{(t-1)} = \mathcal{L}_k(\Psi_k^{\text{loc},(t-1)}) - \mathcal{L}_k^*$. Recall our proposed personalized aggregation scheme in Eq. (8), it can be rewritten as:

$$\Psi_k^{(t)} = \sum_{l=1}^K w_{kl}^{(t-1)} \Psi_l^{\text{loc},(t-1)}. \quad (44)$$

Since each new parameter is a convex combination Eq. (44), based on the Jensen's inequality and $\mathcal{L}$ -smoothness assumption in Eq. (33), the following inequality holds:

$$\mathcal{L}_k(\Psi_k^{(t)}) = \sum_{l=1}^K w_{kl}^{(t-1)} \mathcal{L}_k(\Psi_l^{\text{loc},(t-1)}). \quad (45)$$

²Here, we slightly abuse the notation μ , which was previously introduced in Section C.1.1 with a different meaning.

Let $\mathbf{p} = [p_1, \dots, p_K]^\top$, Eq. (45) subtracts $\mathcal{L}_k^*$ and multiply by p_k , and sum over k , we can reach:

$$\mathcal{L}(\Psi^{(t)}) - \mathcal{L}^* \leq \mathbf{p}^\top \Omega^{(t-1)} \mathbf{f}^{(t-1)}. \quad (46)$$

Let the global average gap as $\bar{f}^{(t-1)} := \mathbf{p}^\top \mathbf{f}^{(t-1)}$ and $\mathbf{r}^{(t-1)} := \mathbf{f}^{(t-1)} - \bar{f}^{(t-1)} \mathbf{1}$. Since row-stochasticity implies $\Omega^{(t-1)} \mathbf{1} = \mathbf{1}$, we have:

$$\mathbf{p}^\top \Omega^{(t-1)} \mathbf{f}^{(t-1)} = \bar{f}^{(t-1)} + \mathbf{p}^\top \Omega^{(t-1)} \mathbf{r}^{(t-1)}. \quad (47)$$

Hence:

$$\mathcal{L}(\Psi^{(t)}) - \mathcal{L}^* \leq \bar{f}^{(t-1)} + \mathbf{p}^\top \mathbf{V}^{(t-1)} \mathbf{r}^{(t-1)}, \quad \mathbf{V}^{(t-1)} := \Omega^{(t-1)} - \frac{1}{K} \mathbf{1} \mathbf{1}^\top. \quad (48)$$

Note that $\mathbf{p}^\top \mathbf{r}^{(t-1)} = 0$ by definition of $\bar{f}^{(t-1)}$. Applying Cauchy-Schwarz we have:

$$\left| \mathbf{p}^\top \mathbf{V}^{(t-1)} \mathbf{r}^{(t-1)} \right| \leq \left\| \mathbf{p}^\top \mathbf{V}^{(t-1)} \right\|_2 \left\| \mathbf{r}^{(t-1)} \right\|_2. \quad (49)$$

Due to the assumption that $\|\mathbf{V}^{(t-1)}\|_2 \leq \rho$ and $\|\mathbf{p}\|_2 \leq 1$ which is a probability vector, we have:

$$\left| \mathbf{p}^\top \mathbf{V}^{(t-1)} \mathbf{r}^{(t-1)} \right| \leq \rho \left\| \mathbf{r}^{(t-1)} \right\|_2. \quad (50)$$

Next bound $\|\mathbf{r}^{(t-1)}\|_2$ by the mean gap as:

$$\begin{aligned} \left\| \mathbf{r}^{(t-1)} \right\|_2^2 &= \sum_k \left(f_k^{(t-1)} - \bar{f}^{(t-1)} \right)^2 \\ &\leq \sum_k \left(f_k^{(t-1)} \right)^2 \\ &\leq \left(\max_k f_k^{(t-1)} \right) \sum_k f_k^{(t-1)} \\ &= \frac{\max_k f_k^{(t-1)}}{\min_k p_k} \left(\mathbf{p}^\top \mathbf{f}^{(t-1)} \right) \\ &\leq \frac{1}{\min_k p_k}. \end{aligned} \quad (51)$$

Let $c := 1/\sqrt{\min_k p_k} \leq \sqrt{K}$, we can combine Eq. (50) and Eq. (51) to get the following inequality:

$$\left| \mathbf{p}^\top \mathbf{V}^{(t-1)} \mathbf{r}^{(t-1)} \right| \leq \rho c \sqrt{\bar{f}^{(t-1)}}. \quad (52)$$

Then by squaring both sides of Eq. (52) and use $\sqrt{\bar{f}} \leq 1 + \bar{f}$, we have:

$$\left| \mathbf{p}^\top \mathbf{V}^{(t-1)} \mathbf{r}^{(t-1)} \right| \leq \rho^2 c^2 \left(1 + \bar{f}^{(t-1)} \right) \leq \frac{2\rho^2 c^2}{1 - \rho} \bar{f}^{(t-1)}, \quad (53)$$

where the last inequality employs $\bar{f}^{(t-1)} \leq (1 - \rho)^{-1} \bar{f}^{(t-1)}$ which is trivial for $0 < \rho < 1$. Taking expectations, we can substitute Eq. (53) in to Eq. (48) to obtain:

$$\mathbb{E} \left[\mathcal{L}(\Psi^{(t)}) - \mathcal{L}^* \right] \leq \left(1 + \frac{2\rho^2}{1 - \rho} \right) \mathbb{E} \left[\bar{f}^{(t-1)} \right]. \quad (54)$$

Given Eq. (43) and $\bar{f}^{(t-1)} = \sum_k p_k \mathbb{E} \left[f_k^{(t-1)} \right]$, we can bound $\mathbb{E} \left[f_k^{(t-1)} \right]$ by:

$$\mathbb{E} \left[f_k^{(t-1)} \right] \leq (1 - \eta\mu)^Q \left(\mathcal{L}_k(\Psi_k^{(t-1)}) - \mathcal{L}_k^* \right) + \frac{\eta \mathcal{L} \zeta^2}{2\mu}. \quad (55)$$

By taking a weighted sum of the above formula over all K clients with weights p_k , we obtain:

$$\mathbb{E} \left[\bar{f}^{(t-1)} \right] \leq (1 - \eta\mu)^Q \left(\mathcal{L}(\Psi^{(t-1)}) - \mathcal{L}^* \right) + \frac{\eta \mathcal{L} \zeta^2}{2\mu}. \quad (56)$$

One-round contraction By plugging Eq. (56) into Eq. (54), we have:

$$\mathbb{E} \left[\mathcal{L} \left(\Psi^{(t)} \right) - \mathcal{L}^* \right] \leq (1 - \eta\mu)^Q \left(1 + \frac{2\rho^2}{1 - \rho} \right) \left(\mathcal{L} \left(\Psi^{(t-1)} \right) - \mathcal{L}^* \right) + \frac{\eta\mathcal{L}\zeta^2}{2\mu} \left(1 + \frac{2\rho^2}{1 - \rho} \right). \quad (57)$$

Since $1 + \frac{2\rho^2}{1 - \rho} \leq 1 + \frac{2\rho}{1 - \rho} = \frac{1}{1 - \rho}$ and $\rho < 1$, the following inequality holds:

$$\mathbb{E} \left[\mathcal{L} \left(\Psi^{(t)} \right) - \mathcal{L}^* \right] \leq (1 - \eta\mu)^Q \left(\mathcal{L} \left(\Psi^{(t-1)} \right) - \mathcal{L}^* \right) + \frac{\eta\mathcal{L}\zeta^2}{2\mu} + \frac{2\eta\mathcal{L}\rho^2}{\mu(1 - \rho)^2}, \quad (58)$$

where the last term absorbs the factor from $(1 - \rho)^{-1}$.

Across T communication rounds By setting $\gamma := (1 - \eta\mu)^Q$ where $0 < \gamma < 1$, we can unroll Eq. (58) as:

$$\mathbb{E} \left[\mathcal{L} \left(\Psi^{(T)} \right) - \mathcal{L}^* \right] \leq \gamma^T \left(\mathcal{L} \left(\Psi^{(0)} \right) - \mathcal{L}^* \right) + \frac{\eta\mathcal{L}\zeta^2}{2\mu} \sum_{t=0}^{T-1} \gamma^t + \frac{2\eta\mathcal{L}\rho^2}{\mu(1 - \rho)^2} \sum_{t=0}^{T-1} \gamma^t. \quad (59)$$

Since the geometric sums satisfy $\sum_{t=0}^{T-1} \gamma^t \leq \frac{1}{1 - \gamma}$, while $1 - \gamma = 1 - (1 - \eta\mu)^Q \geq \eta\mu$, we have:

$$\sum_{t=0}^{T-1} \gamma^t \leq \frac{1}{\eta\mu}. \quad (60)$$

By inserting the above inequality into Eq. (59) and simplify, we can easily get:

$$\mathbb{E} \left[\mathcal{L} \left(\Psi^{(T)} \right) - \mathcal{L}^* \right] \leq \gamma^T \left(\mathcal{L} \left(\Psi^{(0)} \right) - \mathcal{L}^* \right) + \frac{\eta\mathcal{L}\zeta^2}{2\mu} + \frac{2\eta\mathcal{L}\rho^2}{\mu(1 - \rho)^2}. \quad (61)$$

Recovering $\gamma^T = (1 - \eta\mu)^{QT}$ gives exactly Eq. (12) in Theorem 3.3, which concludes the proof.

D Experimental Details

D.1 Dataset Statistics

Table 4 summarizes the statistics of the datasets used in our experiments. It includes the total number of nodes, edges, node classes, and feature dimensions for each dataset. Specifically, we use four citation graph datasets (Cora, CiteSeer, Pubmed, and ogbn-arxiv) and two product co-purchase graph datasets (Amazon-Computer and Amazon-Photo).

Table 4: Dataset statistics

Datasets	Nodes	Edges	Classes	Features
Cora	2,708	5,429	7	1,433
CiteSeer	3,327	4,732	6	3,703
Pubmed	19,717	44,324	3	500
Amazon-Computer	13,752	491,722	10	767
Amazon-Photo	7,650	238,162	8	745
ogbn-arxiv	169,343	2,315,598	40	128

D.2 Implementation Details

Following the standard FL settings, in our FedAux, both the local client models and the global server model adopt the same backbone architecture. We employ MaskedGCN [3] to generate node embeddings and sweep the number of GCN layers over $L \in \{1, 2, 3\}$. The hidden dimension is selected from $d' \in \{64, 128, 256\}$, and dropout probabilities are set to 0.5. The auxiliary projection vector $\mathbf{a}$ is initialized from a Gaussian distribution in $\mathbb{R}^{d'}$. The similarity-temperature parameter α is set to 10, and the bandwidth σ is fixed to 1. For the FL schedule, we run $T = 100$ communication rounds with $Q = 1$ local epoch on the smaller citation datasets (Cora, CiteSeer, Pubmed). On all other datasets, we set the total number of rounds to $T = 200$ and the number of local epochs per round to $Q = 2$. All experiments are executed on a workstation equipped with an NVIDIA Tesla V100 SXM2 GPU (32 GB) running CUDA 12.4.

Table 5: The degree of non-IIDness.

Non-IIDness	Cora			CiteSeer			Pubmed		
	5 Clients	10 Clients	20 Clients	5 Clients	10 Clients	20 Clients	5 Clients	10 Clients	20 Clients
ξ	0.2667	0.3092	0.3760	0.1848	0.2292	0.2572	0.1316	0.1500	0.1725
Non-IIDness	Amazon-Computer			Amazon-Photo			ogbn-arxiv		
	5 Clients	10 Clients	20 Clients	5 Clients	10 Clients	20 Clients	5 Clients	10 Clients	20 Clients
ξ	0.2774	0.3582	0.3931	0.3600	0.4314	0.4840	0.3398	0.3668	0.4307

D.3 Quantifying Non-IIDness in Federated Graph Datasets

To compare how much statistical heterogeneity (i.e., non-IIDness) each dataset induces under a given partition scheme, it is useful to measure how far the local data distribution at each client deviates from the global distribution and how dispersed local distributions are from one another. Below are three complementary, fully formalized metrics that can be computed once the graph has been split into K client subgraphs $\{G_1, \dots, G_K\}$.

Label-distribution divergence Let $P(y)$ be the global class prior and $P_k(y)$ the class prior in client G_k . We use the average Jensen–Shannon (JS) divergence to measure the gap between each local label prior and the global label prior as follows:

$$\text{JSD} = \frac{1}{K} \sum_{k=1}^K \frac{1}{2} [\text{KL}(P_k \| R_k) + \text{KL}(P \| R_k)], \quad (62)$$

where $R_k = \frac{1}{2}(P_k + P)$ denotes the mid-point distribution between P_k and P with $R_k(y \in Y) = \frac{1}{2} [P_k(y) + P(y)]$. Pinsker’s inequality [6] gives that $\text{JSD} \in [0, \log 2]$. A small JSD indicates that each client’s label distribution closely matches the global prior, so the partition is effectively IID. As the JSD increases, local class proportions deviate more sharply from the global mixture, making individual clients progressively class-specific and therefore increasingly subject to statistical non-IIDness.

Subgraph-distribution discrepancy Label skew alone may underestimate heterogeneity when covariate shift is strong. To quantify covariate-shift–induced heterogeneity, we measure how far the embedding distribution of each client deviates from that of every other client in an embedding space that reflects graph structure and attributes. We first obtain node embeddings for all nodes with a simple neighbor aggregation $\mathbf{Z}_k = \mathbf{A}_k \mathbf{X}_k = \{z_{k,i}\}_{i=1}^{N_k}$ in each client. Let $\mathbf{Z}_k$ be the empirical distribution of embeddings held by client G_k . We define the graph-distribution discrepancy of a K -client partition as the mean pair-wise maximum mean discrepancy (MMD) as:

$$\text{MMD} = \frac{2}{K(K-1)} \sum_{1 \leq k < l \leq K} \left\| \frac{1}{|V_k|} \sum_{v_i \in V_k} \phi(z_{k,i}) - \frac{1}{|V_l|} \sum_{v_j \in V_l} \phi(z_{l,j}) \right\|^2, \quad (63)$$

where $\phi(\cdot)$ is the canonical feature map of a Gaussian RBF kernel with the bandwidth fixed with the median pairwise distance to ensure comparability across datasets. Eq. (63) evaluates to zero when all clients share an identical embedding distribution (IID) and increases monotonically with covariate divergence.

Based on the above two perspectives, we can quantify the degree of non-IIDness by summing JSD and MMD as $\xi = \text{JSD} + \text{MMD}$, where a higher ξ indicates a higher degree of non-IIDness. We show the degree of non-IIDness of all datasets under different numbers of clients in Table 5.

D.4 Synthetic Graph for Client Similarity Estimation

We first generate an SBM graph with 3000 nodes that are uniformly divided into $K = 20$ equal-sized blocks as clients. Inter-client edges are added with probability $P^{\text{inter}} = 0.02$. To inject structural non-IIDness, the 20 clients are grouped into five super-clusters $\{\mathcal{G}_i\}_{i=1}^5$ (four clients per group). For every client belonging to $\mathcal{G}_i$, we draw intra-client edges with probability $P_i^{\text{intra}} = 0.15 \times i$, so clients

Table 6: Ablation studies on the federated node classification task under 10 clients.

Baseline	Cora	Pubmed	Amazon-Computer	ogbn-arxiv
(i) FedAux _{hard}	80.29 $\pm$ 0.71	83.06 $\pm$ 0.35	88.10 $\pm$ 1.02	66.53 $\pm$ 0.12
(ii) FedAvg _{mask}	78.98 $\pm$ 0.54	83.97 $\pm$ 0.51	84.31 $\pm$ 0.60	65.09 $\pm$ 0.07
FedAux	82.05$\pm$0.71	85.43$\pm$0.29	89.92$\pm$0.15	68.50$\pm$0.27

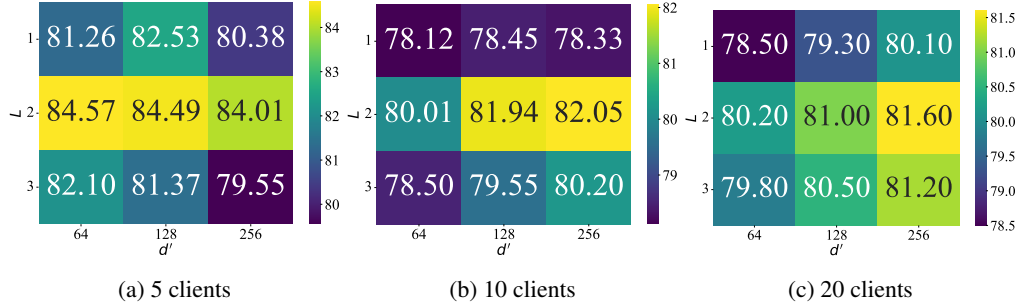


Figure 7: Classification accuracy (%) for different GCN configurations.

within the same group are structurally IID, while clients across groups are non-IID. To inject label and feature non-IIDness, we assign 5 labeled classes to all nodes. Nodes owned by the group $\mathcal{G}_i$ receive label $i - 1$ with high probability 0.8, and the remaining 0.2 mass is distributed uniformly over the other four labels. Each node’s feature vector is the one-hot encoding of its label. In this way, clients in the same group are IID in both label and feature space, while clients from different groups exhibit pronounced distributional shifts. This controlled setting allows us to test whether the learned APVs can cluster clients that are genuinely similar. We directly compute the similarity between clients’ mean-pooling embeddings without considering privacy in Fig. 3a as the ground-truth similarity. Fig. 3b-3d are similarities computed by APVs, learnable weights of the readout layer, and functional embedding [3].

E More Experiments

E.1 Ablation Study

We conduct an ablation study in Table 6 to validate our motivation and design, with the following two ablations: (i) replacing our proposed continuous aggregation scheme over the α_k -space defined in Eq. (6) with the Conv1D operation applied on the hard-sorted embeddings as introduced by [18], yielding a variant FedAux_{hard}, and (ii) removing our server-side APV-based personalized aggregation, instead using simple averaging to aggregate local models into a global model, leading to the baseline FedAvg_{mask}. Note that the FedAvg_{mask} variant differs from the standard FedAvg used in Table 1, where FedAvg_{mask} adopts MaskedGCN [3] as the GNN backbone, whereas the standard FedAvg utilizes a conventional GCN architecture. Compared with FedAux_{hard}, our FedAux consistently obtains higher accuracy. This empirical gain confirms our claim that although the hard-sorting scheme proposed by [18] does allow gradient flow to optimize the APV, the underlying discrete permutation remains non-differentiable and therefore restricts the capacity of the APV to adapt. By using our proposed continuous aggregation with a fully differentiable kernel operator, FedAux enables smoother gradient flow, allowing the APV and the local GNN to co-evolve optimally. Besides, FedAux and FedAvg_{mask} both adopt MaskedGCN as backbone, while FedAux personalized aggregates local models based on the APV similarity, while FedAvg_{mask} aggregates local models to a single global model. Results show that FedAux outperforms FedAvg_{mask} on all datasets, which demonstrates that exploiting APV-based personalized aggregation allows the federation to respect non-IID data distributions and learn more effective client-specific models.

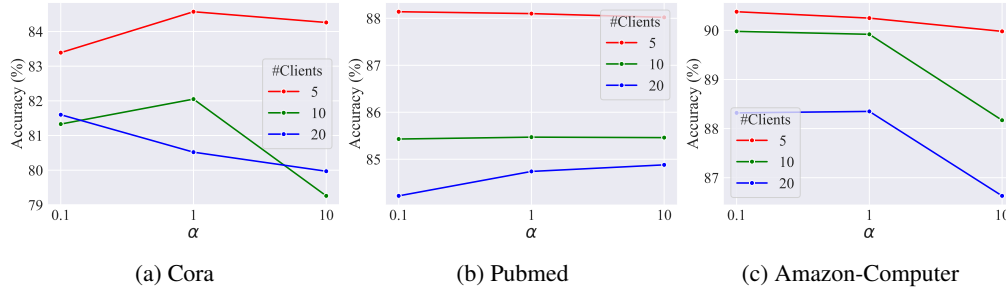


Figure 8: Sensitivity of FedAux on the similarity temperature parameter α .

E.2 Hyperparameter Analysis

Impact of Model Depth and Hidden Dimension In our model, the number of GCN layers is selected from $L \in \{1, 2, 3\}$, and the dimension of the hidden layers is selected from $d' \in \{64, 128, 256\}$. In Fig. 7, we show all the hyperparameter combinations on the Cora dataset for different client counts. It is evident that FedAux consistently achieves the highest performance with $L = 2$ across all federated settings. For the hidden dimension, when the number of clients is not large, FedAux requires a relatively small hidden dimension, while with 20 clients, a hidden dimension of 256 yields the best results.

Impact of Similarity Temperature α In Eq. (7), the similarity temperature parameter α is introduced to modulate the sharpness of the similarity distribution. To evaluate the sensitivity of FedAux to this hyperparameter, we test $\alpha \in \{0.1, 1, 10\}$ across varying numbers of clients and report the resulting accuracy in Fig. 8. The results show that while some configurations achieve optimal accuracy at different values of α , for example, $\alpha = 0.1$ is optimal for 20 clients on Cora and $\alpha = 10$ is optimal for 20 clients on Pubmed, directly setting $\alpha = 1$ consistently provides satisfactory performance across different datasets and client counts.

F Limitations and Future Work

The main limitation of our work is that while the kernel-based continuous aggregation scheme lets each client learn a fully differentiable, data-driven ordering on the APV, it requires evaluating the pairwise kernel for every node pair in the local client subgraph. Thus, the per-client computational cost scales quadratically, $\mathcal{O}(N_k^2)$, which is acceptable for small and medium subgraphs but can dominate run-time on very large clients. Our future work focuses on addressing this scalability bottleneck. One promising direction is to approximate the dense kernel with low-rank factorizations, such as Nyström approximation, so that complexity scales linearly in N_k with controllable error. We stress that our core contribution is a privacy-preserving mechanism for personalized clustering clients via the APV, and reducing local computational load is a complementary line of future research.