Beyond the Federation: Topology-aware Federated Learning for Generalization to Unseen Clients

Mengmeng Ma[1] **Tang Li**[1] **Xi Peng**[1]

Abstract

Federated Learning is widely employed to tackle distributed sensitive data. Existing methods primarily focus on addressing in-federation data heterogeneity. However, we observed that they suffer from significant performance degradation when applied to unseen clients for out-of-federation (OOF) generalization. The recent attempts to address generalization to unseen clients generally struggle to scale up to large-scale distributed settings due to high communication or computation costs. Moreover, methods that scale well often demonstrate poor generalization capability. To achieve OOF-resiliency in a scalable manner, we propose Topology-aware Federated Learning (TFL) that leverages client topology - a graph representing client relationships - to effectively train robust models against OOF data. We formulate a novel optimization problem for TFL, consisting of two key modules: Client Topology Learning, which infers the client relationships in a privacy-preserving manner, and Learning on Client Topology, which leverages the learned topology to identify influential clients and harness this information into the FL optimization process to efficiently build robust models. Empirical evaluation on a variety of real-world datasets verifies TFL's superior OOF robustness and scalability.

1. Introduction

Federated Learning (FL) has emerged as a promising solution for handling distributed sensitive data, enabling multi-institutional collaboration by distributing model training to data owners and aggregating results on a centralized server (McMahan et al., 2017). This data-decentralized
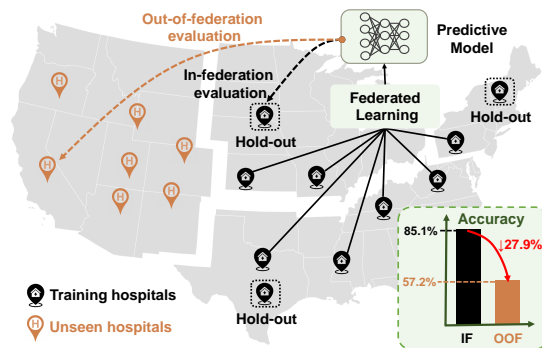

Figure 1. Our empirical evaluation of patient mortality prediction using Federated Learning on distributed Healthcare dataset (eICU (Pollard et al., 2018)). We observed that *a model exhibiting high accuracy with in-federation (IF) data can fail catastrophically when presented with out-of-federation (OOF) data.*

approach harnesses the collective intelligence of all participating nodes to build a model that is potentially more robust and generalizable (Sheller et al., 2020). Existing robust FL methods (Li et al., 2020b; Deng et al., 2020) primarily focus on learning a global model with good average or worst-case performance, addressing *in-federation (IF)* data heterogeneity. However, these methods can fail catastrophically on *out-of-federation (OOF)* clients, *i.e.,* clients outside the collaborative federation. The OOF clients pose significant generalization challenges, as FL models may encounter *unseen* distributions outside their training space (Pati et al., 2022a). Our empirical study shows that existing methods suffer from significant performance degradation when applied to unseen clients for OOF generalization (see Figure 1).

There have been recent attempts to address the challenge of unseen clients through client augmentation and client alignment. However, these approaches encounter difficulties in large-scale distributed environments. Client augmentation methods (Liu et al., 2021) often necessitate intensive client-server communication, leading to a notable *communication bottleneck* that hinders scalability (Zhou et al., 2023). Client alignment methods (Zhang et al., 2021), which operate by aligning client distributions through adversarial training, are less communication costly but introduce substantial *computation burdens* due to complex local training process (Bai et al., 2023). While using a fixed reference distribution for alignment (Nguyen et al., 2022) offers computational effi-
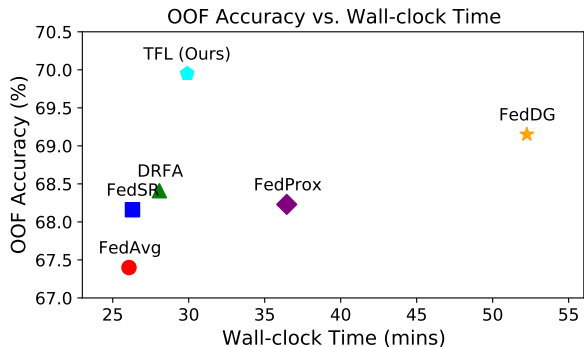
---

[1]DeepREAL Lab, Department of Computer & Information Sciences, University of Delaware. Correspondence to: Mengmeng Ma <mengma@udel.edu>.

Figure 2. Accuracy vs. wall-clock time on PACS dataset (Li et al., 2017). Wall-clock time is used as a holistic evaluation of scalability (communication and computation costs). *We see a tradeoff between OOF robustness and scalability in existing methods.*

ciency, it has shown *limited generalization capabilities to unseen clients* (Bai et al., 2023). Our empirical evaluation (Figure 2) reveals the tradeoff between OOF resiliency and scalability in current methods. A question naturally arises: Can we build OOF robust models in a scalable manner?

We propose to trigger OOF-resiliency while being scalable by leveraging *client topology*. Client topology, a graph representing client relationships, allows for using graph mining techniques (Saxena & Iyengar, 2020; Lü & Zhou, 2011) to derive insights into client data distributions. It can be used to identify "influential" clients that are representative of the training clients, containing distributions more likely to be encountered in OOF clients. For instance, an influential client could be a regional hospital that aggregates a diverse range of patient data. This hospital's data encapsulates a rich repository of information, mirroring the variety and complexity of data that could be seen in OOF scenarios. Leveraging these influential clients as priority contributors in the training rounds can facilitate the model in learning from the most representative data, thereby potentially enhancing its OOF robustness. On the other hand, by reducing unnecessary communication with non-influential clients, communication costs can be significantly reduced.

Grounded on the concept of client topology, we formulate the problem of generalization to unseen clients as the optimization of a joint objective over the models and the client topology. To solve this optimization problem, we propose Topology-aware Federated Learning, which consists of two steps: 1) Client Topology Learning: Inferring the client topology while respecting data privacy. 2) Learning on Client Topology: Leveraging the learned topology to build a robust model. The first step learn a client topology that reflects client relationships by analyzing pairwise model similarity. In the second step, a robust model is efficiently optimized by harnessing the client's influential information to regularize a distributed robust optimization process.

Our contributions are as follows: Firstly, we introduce the *Topology-aware Federated Learning* (TFL), a scal-able framework designed to enhance FL's out-of-federation (OOF) robustness. TFL utilizes client relationships to develop robust models against OOF data. Secondly, we design an iterative client topology learning and learning on client topology approach to solve TFL. Finally, we have *curated two OOF benchmarks* using real-world healthcare data, offering valuable testbeds for subsequent research. Through extensive experiments on curated and standard benchmarks, we verify TFL's superior OOF-resiliency and scalability.

2. Preliminaries

We start by presenting the average-case formulation of FL and subsequently introduce a fundamental robust optimization framework for FL (worst-case formulation).

Federated learning (Average-case formulation). The standard FL involves collaboratively training a global model leveraging data distributed at K clients. Each client k ($1 \leq k \leq K$) has its own data distribution $D_k(x, y)$, where $x \in \mathcal{X}$ is the input and $y \in \mathcal{Y}$ is the label, and a dataset with n_k data points: $\widehat{D}_k = \{(x_k^n, y_k^n)\}_{n=1}^{n_k}$. Local data distributions $\{D_k\}_{k=1}^K$ could be the same or different across the clients. FL aims to learn a global model θ by minimizing the following objective function:

$$\min_{\theta \in \Theta} F(\theta), \text{ where } F(\theta) := \sum_{k=1}^{K} p_k f_k(\theta), \qquad (1)$$

where $f_k(\theta)$ is the local objective function of client k. The local objective function is often defined as the empirical risk over local data, *i.e.*, $f_k(\theta) = \mathbb{E}_{(x,y) \sim \widehat{D}_k}[\ell(\theta; x, y)] = \frac{1}{n_k} \sum_{n=1}^{n_k} \ell(\theta; x_k^n, y_k^n)$. The term p_k ($p_k \geq 0$ and $\sum_k p_k = 1$) specifies the relative importance of each client, with two settings being $p_k = \frac{1}{N}$ or $p_k = \frac{n_k}{N}$, where $N = \sum_k n_k$ is the total number of samples.

Distributionally robust federated learning (Worst-case formulation). While Equation 1 can build a global model with good *average performance* on in-federation clients, it may not necessarily guarantee good performance in the presence of heterogeneous data. In the real world, data could be statistically heterogeneous due to different data acquisition protocols or various local demographics (Rieke et al., 2020). Thus, the local data distributions may deviate significantly from the average distribution, implying that an "average" global model can fail catastrophically under distributional drift. To tackle this issue, distributionally robust optimization (Staib & Jegelka, 2019) has been adapted to FL, resulting in distributionally robust federated learning (Deng et al., 2020; Reisizadeh et al., 2020). The formulation of this new problem is as follows:

$$\min_{\theta \in \Theta} \max_{\boldsymbol{\lambda} \in \Delta_K} F(\theta, \boldsymbol{\lambda}) := \sum_{k=1}^{K} \lambda_k f_k(\theta), \qquad (2)$$

where $\boldsymbol{\lambda}$ is the global weight for each local loss function
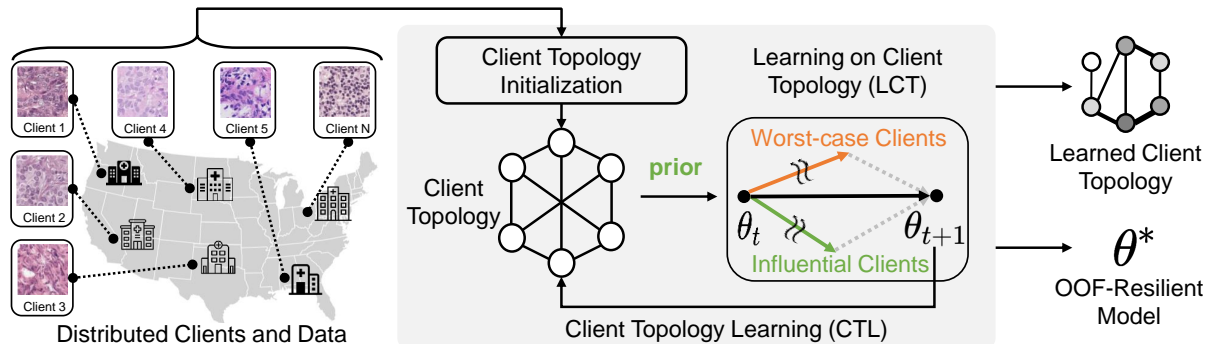
Figure 3. Overview of Topology-aware Federated Learning (TFL). TFL contains two alternating steps: *client topology learning* (CTL) and *learning on client topology* (LCT). **CTL** learn the client topology that describes the relationships between local clients. We use model weights as node embedding and construct a ϵ-graph by measuring the similarity of node pairs. **LCT** leverage the learned client topology to achieve better OOF robustness. We identify the influential client and then use the influential client as prior knowledge to regularize a distributionally robust optimization framework. In this way, the optimization process can balance the "worst-case" client and the "influential" client to avoid overly pessimistic models with compromised OOF-resiliency.

and Δ_K denotes the $K-1$ probability simplex. Intuitively, Equation 2 minimizes the maximal risk over the combination of empirical local distributions, and therefore the worst-case clients would be prioritized during training.

While this framework has the potential to address distribution shifts (Mohri et al., 2019; Qiao & Peng, 2023; Deng et al., 2020), directly implementing it for OOF resiliency may yield suboptimal models. This approach heavily relies on *worst-case clients*, those with large empirical risks, to develop robust models. However, these clients may not necessarily be the *influential ones* that are representative of clients. In some cases, it is possible that this approach overly focuses on "outlier" clients, clients that are significantly different from most of the training clients, leading to models with limited OOF robustness. Therefore, we argue that, to build optimal OOF-resilient models, the optimization process should focus on not only the worst-case but also the influential clients.

3. Methodology

In this section, we introduce the proposed *Topology-aware Federated Learning* (TFL) framework. TFL aims to leverage client topology to improve the model's OOF robustness. We use graphs as tools to describe the client topology because they provide a flexible way to represent the relationship between distributed clients (Dong et al., 2019; Mateos et al., 2019). Specifically, we model client topology using an undirected graph (Vanhaesebrouck et al., 2017; Ye et al., 2023). In the graph, nodes correspond to clients, and edges reflect client connectivity. Let $\mathcal{G} = (V, E, W)$ denote the client topology, where V is the node set with $|V| = K$, $E \subseteq V \times V$ is the edge set and $W \in \mathbb{R}^{K \times K}$ is the adjacency matrix. An edge between two clients v_k and v_l is represented by $e_{k,l}$ and is associated with a weight $w_{k,l}$.

Optimization problem. Let ϕ denotes graph measures, *e.g,* centrality measure (Saxena & Iyengar, 2020). The function "sim" indicates any similarity function, including but not limited to cosine similarity, ℓ_2, or ℓ_1. γ represents a trade-off hyperparameter. $\mathcal{D}$ symbolizes arbitrary distributional distance measure. As the client topology is often not readily available, we propose to jointly optimize the client topology and the robust model by solving the following problem:

$$\min_{\substack{\theta \in \Theta \\ w \in W}} \max_{\boldsymbol{\lambda} \in \Delta_K} F(\theta, \boldsymbol{\lambda}, W) := \sum_{k=1}^{K} \lambda_k f_k(\theta)$$
$$- \frac{\gamma}{2} \sum_{k \neq l} w_{k,l} \, \text{sim}(v_k, v_l), \quad (3)$$
$$\text{s.t. } \mathcal{D}(\boldsymbol{\lambda} \parallel \boldsymbol{p}) \leq \tau, \text{ where } \boldsymbol{p} = \phi(W).$$

In the proposed objective function, the first term follows the same spirit of Equation 2 to adopt a minimax robust optimization framework. The difference is that it minimizes the risk over not only the worst-case but also the influential clients. The second term is dedicated to learning the client topology by measuring the pair-wise client similarity.

Our formulation stands apart from existing work in two respects. First, it employs client topology to explicitly model the client relationships. Analyzing this topology facilitates the identification of influential clients that are crucial for developing OOF generalization. Second, our formulation enables seamless integration of client topology into the optimization process, ensuring that the model assimilates insights from the most significant clients.

In Equation 3, it is infeasible to simultaneously update both the client topology W and model parameters θ as local clients do not have access to other clients' data. To this end, we propose to solve this problem using an alternating two-step approach: learning the client topology (updating W)

and building an OOF-resilient model with client topology (updating λ and θ). As shown in Figure 3, we term these steps as **Client Topology Learning** (CTL) and **Learning on Client Topology** (LCT), respectively. The following sections will provide further details on these two steps.

3.1. Client Topology Learning

Our goal is to learn the client topology that accurately describes the characteristics of local data, capturing the underlying relationships among clients. Conventional approaches typically adopt similarity-based (Chen et al., 2020a; Franceschi et al., 2019) or diffusion-based (Zhu et al., 2021; Mateos et al., 2019) methods to estimate the graph structure from data. However, most methods require centralizing training data on a single machine, thereby raising privacy concerns. Therefore, the primary challenge lies in learning the client topology while respecting data privacy.

We propose to utilize model weights to learn client topology. Intuitively, when two clients have similar data distributions, their corresponding models should be more similar. It is possible to obtain data distribution relationships with model similarity. While the feasibility of this idea is supported by literature (Yu et al., 2022), we intend to empirically verify whether it holds in our setting. We use four types of similarity measures, including ℓ_1-based, ℓ_2-based, dot product-based, and cosine similarity-based metrics. We conduct experiments on PACS where clients 6 to 10 share similar data distributions. Then, we measure the similarity between client 10 and all other clients. For a clear comparison, we normalized all scores into $[0, 1]$. The results are shown in Figure 4. We observe that clients sharing similar data distributions exhibit significantly higher similarity scores compared to others. Furthermore, these distinct similarity scores remain consistent across various metrics.

In summary, utilizing model weights to learn client topology offers two merits. First, models can be freely shared among clients, addressing privacy issues of client topology learning. Second, since models are trained on local data, their similarity measures the similarity between local data distributions. We formulate client topology learning as follows:

$$\min_{w \in W} - \sum_{k \neq l} w_{k,l} \, \mathrm{sim}(\theta_k, \theta_l) + \|W\|_0, \qquad (4)$$

where θ denotes the model of local clients. This objective ensures that similar clients have large edge weights. Our formulation is aligned with the common network homophily assumption that *edges tend to connect similar nodes* (Newman, 2018). To avoid learning a fully connected graph with trivial edges, we enforced graph sparsity by penalizing the l_0-norm of the adjacency matrix $\|W\|_0$. We implement the sparsity term using an implicit method, *i.e.,* hard thresholding on W to construct ϵ-graph (Zhu et al., 2021). Specifi-
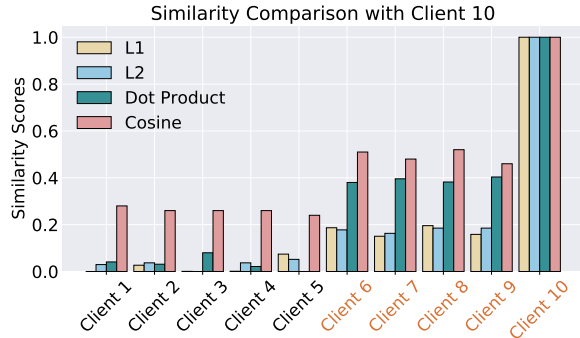

Figure 4. Model similarity under different metrics. Client models are trained using the same algorithm and hyperparameters. Clients 6 to 10 share similar data distributions. We observe that *clients with similar data distribution tend to have more similar models.*

cally, we mask out (*i.e.,* set to zero) those elements in W smaller than a non-negative threshold ϵ.

3.2. Learning on Client Topology

The learned client topology captures the relation among local clients. We aim to leverage such relations to develop TFL for better OOF-resiliency. Recall that, to tackle distribution shift, distributionally robust federated learning (DRFL) assumes that the target distribution lies within an arbitrary mixture of training distributions: $\sum_{k=1}^{K} \lambda_k D_k$. DRFL builds OOF-resilient models by minimizing the *worst-case* risk over an uncertainty set of possible target distribution $Q := \{\sum_{k=1}^{K} \lambda_k D_k \mid \boldsymbol{\lambda} \in \Delta_K\}$. However, DRFL primarily emphasizes the worst-case distribution, potentially ignoring the *influential* ones that are more representative of the training clients. This can lead to overly pessimistic models with compromised OOF resiliency (Hu et al., 2018).

We leverage client topology to construct an uncertainty set that can better approximate the unseen distribution. Our insight is to optimize the model for both the worst-case and influential distributions. The key challenge is how to identify the influential distribution. Our idea is to use graph centrality as the criterion to choose influential distributions. *Graph centrality* is widely used in social network analysis (Newman, 2005) to identify the key person by measuring how much information propagates through each entity. We introduce *client centrality* to identify influential ones, which can be calculated by graph measurements such as degree, closeness, and betweenness. Specifically, we first calculate the centrality of each client in $\mathcal{G}$ as the *topological prior $\boldsymbol{p}$*. Then we use $\boldsymbol{p}$ to constraint the uncertainty set Q by solving the following minimax optimization problem:

$$\min_{\theta \in \Theta} \max_{\boldsymbol{\lambda} \in \Delta_K} F(\theta, \boldsymbol{\lambda}) := \sum_{k=1}^{K} \lambda_k f_k(\theta),$$
$$\text{s.t. } \mathcal{D}(\boldsymbol{\lambda} \parallel \boldsymbol{p}) \leq \tau. \qquad (5)$$

The topological constraint directs the optimization to focus on worst-case and influential clients. Notably, both FedAvg (McMahan et al., 2017) (Equation 1) and DRFL (Equation 2) can be viewed as special cases of TFL: When $\tau = 0$ and the prior $\boldsymbol{p}$ is a uniform distribution, Equation 5 minimizes the average risks over local clients, which is identical to FedAvg; When $\tau \to \infty$, Equation 5 only prioritizes the worst-case clients, resembling the DRFL approach.

The above optimization problem is typically nonconvex, and methods such as SGD cannot guarantee constraint satisfaction (Robey et al., 2021). To tackle this issue, we leverage the Lagrange multiplier and KKT conditions (Boyd et al., 2004) to convert it into unconstrained optimization:

$$\min_{\theta \in \Theta} \max_{\boldsymbol{\lambda} \in \Delta_K} F(\theta, \boldsymbol{\lambda}, \boldsymbol{p}) := \sum_{k=1}^{K} \lambda_k f_k(\theta) - q\mathcal{D}(\boldsymbol{\lambda} \parallel \boldsymbol{p}), \quad (6)$$

where q is the dual variable. We solve the primal-dual problem by alternating between gradient descent and ascent:

$$\begin{aligned} \theta^{t+1} &= \theta^t - \eta_\theta^t \nabla_\theta F(\theta, \boldsymbol{\lambda}), \\ \boldsymbol{\lambda}^{t+1} &= \mathcal{P}_{\Delta_K}(\boldsymbol{\lambda}^t + \eta_\lambda^t \nabla_\lambda F(\theta, \boldsymbol{\lambda})), \end{aligned} \quad (7)$$

where η^t is the step size. $\mathcal{P}_{\Delta_K}(\boldsymbol{\lambda})$ projects $\boldsymbol{\lambda}$ onto the simplex Δ_K for regularization.

Influential client. We identify the influential clients by calculating betweenness centrality. Betweenness centrality measures how often a node is on the shortest path between two other nodes in the graph $\mathcal{G}$. It has been revealed that the high betweenness centrality nodes have more control over the graph as more information will pass through them (White & Borgatti, 1994). The betweenness centrality of client k is given by the expression of $c_k = \sum_{s \neq k \neq t \in [K]} \frac{\sigma_{st}(k)}{\sigma_{st}}$, where σ_{st} is the total number of shortest path from node s to node t ((s, t)-paths) and $\sigma_{st}(k)$ is the number of (s, t)-paths that pass through node k. Then we apply softmax to normalize client centrality c_k to obtain client topological prior $p_k = \exp(c_k)/\sum_{k=1}^{K} \exp(c_k)$.

Discussion. Handling high-dimensional models. Large models are getting more attention in FL (Zhuang et al., 2023). Its high dimensionality might raise computational concerns when calculating model similarity. For better computation efficiency, we can leverage only partial model parameters, *e.g.,* the last few layers. We empirically show that using partial parameters does not significantly affect the OOF performance (see results in Table 7). **Client topology learning in cross-device settings**. Client topology learning incurs $\mathcal{O}(N^2)$ computation complexity for N clients. This quadratic complexity could be prohibitively expensive in cross-device FL that involves thousands or millions of clients. We argue that computation costs can be significantly reduced via client clustering (Sattler et al., 2020). Our empirical evaluation (see results in Table 6) shows that client clustering significantly reduces computation costs.

4. Experiments

4.1. Datasets and Baselines

Datasets. TFL is evaluated on our curated real-world datasets (①eICU, ②FeTS, ③TPT-48) and standard benchmarks (④CIFAR-10/-100, ⑤PACS), spanning a wide range of tasks including classification, regression, and segmentation. Evaluations covers both out-of-federation (datasets ①-③, ⑤) and in-federation (datasets ③-④) scenarios. Below are the descriptions of the real-world datasets. For detailed descriptions of all datasets, including ④CIFAR-10/-100 and ⑤PACS, please refer to Supplementary A.

① **eICU** (Pollard et al., 2018) is a large-scale multi-center critical care database. It contains high granularity critical care data for over $200,000$ patients admitted to 208 hospitals across the United States. We follow (Huang et al., 2019) to pre-process the raw data. The resultant dataset comprises $19,000$ patients over 72 hospitals. We further split the patients into four groups based on census regions and provided the protocol for OOF evaluation.

② **FeTS** (Pati et al., 2022b) is a multi-institutional medical imaging dataset. It comprises clinically acquired MRI scans of glioma. We use a subset of the original data, comprising 358 subjects from 21 distinct global institutions, and provide a protocol for OOF evaluation. The task is brain tumor segmentation and the evaluation metric is Dice Similarity Coefficient (DSC $\uparrow$).

③ **TPT-48** (Vose et al., 2014) contains the monthly average temperature for the 48 contiguous states in the US from 2008 to 2019. The task is to predict the next six months' temperature given the first six months' data. We consider two tasks: (1) E(24) $\to$ W(24): Using the 24 eastern states as IF clients and the 24 western states as OOF clients; (2) N(24) $\to$ S(24): Using the 24 northern states as IF clients and the 24 southern states as OOF clients. The evaluation metric is Mean Squared Error (MSE $\downarrow$).

Baselines. We compare with FedAvg (McMahan et al., 2017) and FedProx (Li et al., 2020b), the most referenced baselines in the literature. DRFA (Deng et al., 2020) is the latest work that adopts the federated distributionally robust optimization framework. FedSR (Nguyen et al., 2022) is the most recent work that tackles FL's generalization to unseen clients. We did not compare with FedDG (Liu et al., 2021) as it requires the sharing of data in the frequency space with each other. This can be viewed as a form of data leakage (Bai et al., 2023; Nguyen et al., 2022). We provide implementation details on model architecture and hyperparameters in Supplementary D.

4.2. Evaluation on OOF-resiliency

Takeaway 1: Learning on client topology improves OOF robustness. We evaluate our method on different datasets

Table 1. Accuracy on the PACS. We conduct experiments using *leave-one-domain-out*, meaning each domain serves as the evaluation domain in turn. Existing methods typically consider each domain as an individual client (Liu et al., 2021; Nguyen et al., 2022). To simulate a large-scale setting, we further divide each domain into 5 subsets and treat each subset as a separate client. The total number of clients is 20. The reported numbers are from three independent runs. *Our method outperformed others across all tested settings.*

Models		Backbone	PACS				
			A	C	P	S	Average
Centralized	DGER (Zhao et al., 2020)	ResNet18	80.70	76.40	96.65	71.77	81.38
Methods	DIRT-GAN (Nguyen et al., 2021)	ResNet18	82.56	76.37	95.65	79.89	83.62
	FedAvg	ResNet18	55.83±0.31	61.37±0.66	77.87±0.61	74.53±0.18	67.40
Federated	FedProx	ResNet18	56.84±0.88	62.56±0.87	78.33±0.46	75.17±0.61	68.23
Learning	DRFA	ResNet18	56.59±0.34	62.87±0.22	78.63±0.77	75.55±0.42	68.41
Methods	FedSR	ResNet18	57.56±0.95	61.91±0.35	78.42±0.19	74.73±0.27	68.16
	TFL (Ours)	ResNet18	**59.05**±0.69	**64.46**±0.21	**79.35**±0.61	**76.93**±0.39	**69.95**

Table 2. Best ROC-AUCs, corresponding communication round, and the total number of communicated parameters (donate as c-params. (M) in the table) on eICU. This dataset comprises EHRs collected from a diverse range of 72 hospitals across the United States. We trained our model using data from 58 hospitals located in the MIDWEST, NORTHEAST, and SOUTH regions. We evaluated the performance of the global model on an independent set of 14 hospitals from the WEST. The reported numbers are from three independent runs. *Our approach achieves the best OOF performance with minimal communication rounds and the number of communicated parameters.*

Centralized Method	Federated Learning Methods														
ERM	FedAvg			FedProx			DRFA			FedSR			TFL (Ours)		
	ROC-AUC	round	c-params.	ROC-AUC	round	c-params.	ROC-AUC	round	c-params.	ROC-AUC	round	c-params.	ROC-AUC	round	c-params.
67.04±1.88	57.18±0.03	6	15.66M	57.21±0.01	6	15.66M	57.20±0.09	2	10.44M	57.25±0.03	8	20.88M	**58.41**±0.06	2	10.44M

and summarize the results in Tables 1, 2, and 3. We make the following observation: our method improves the model's OOF-resiliency on both standard and real-world datasets. From Table 1, our method performs 2.25% better than the federated robust optimization method (DRFA) and 2.63% better than the federated domain generalization method (FedSR). From Table 2, our method performs 2.11% better than DRFA and 2.03% better than FedSR. Lastly, from Table 3, our method performs 1.98% better than DRFA and 3.00% better than FedSR. Overall, our method shows consistently superior OOF robustness than state-of-the-art across all the evaluated datasets. We also visualize the segmentation results of FeTS in Figure 5. Our method delivers high-quality segmentation results, suggesting increased reliability for real-world healthcare applications that have to contend with diverse local demographics.

Visualization of client topology. We also visualize the learned client topology of the eICU dataset in Figure 6. We observe that the learned client topology helps to identify the "influential" clients. The figure indicates that the majority of important clients are from the MIDWEST and SOUTH, with no influential clients from the NORTHEAST.

4.3. Evaluation on Scalability

Takeaway 2: TFL is scalable for OOF generalization. We investigate whether TFL will significantly increase the communication or computation overhead, potentially affecting scalability in large-scale settings. We present ROC AUC, communication rounds, and transmitted parameters for the

Table 3. DSC (↑) score on FeTS. This dataset contains tumor images from 21 institutions worldwide. We conduct training on 15 institutions and evaluate the model on the remaining 5. The reported numbers are from three independent runs. *Our method delivers the best OOF robustness.*

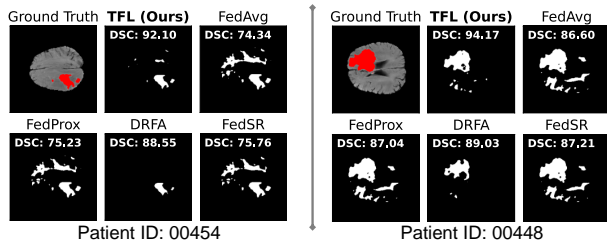| Centralized Method	Federated Learning Methods				
ERM	FedAvg	FedProx	DRFA	FedSR	TFL (Ours)
83.14±0.98	71.45±0.05	71.15±0.04	72.12±1.03	72.85±0.05	**74.29**±0.15

Figure 5. Qualitative results on unseen patients from FeTS dataset. We also show the DSC (↑) score. *Our method consistently demonstrates superior OOF robustness under diverse local demographics.* Additional visualizations can be found in Supplementary A.

eICU dataset in Table 2 and Figure 7. Our method achieves the highest ROC AUC score while requiring the fewest communication rounds and the total number of communicated parameters, thereby indicating its superior communication efficiency. Notably, DRFA also achieves its peak performance within the same number of communication rounds as our method. This is primarily attributable to the fact that both DRFA and our approach utilize a distributionally robust optimization framework. By minimizing the worst-case combination of local client risks, the model converges faster
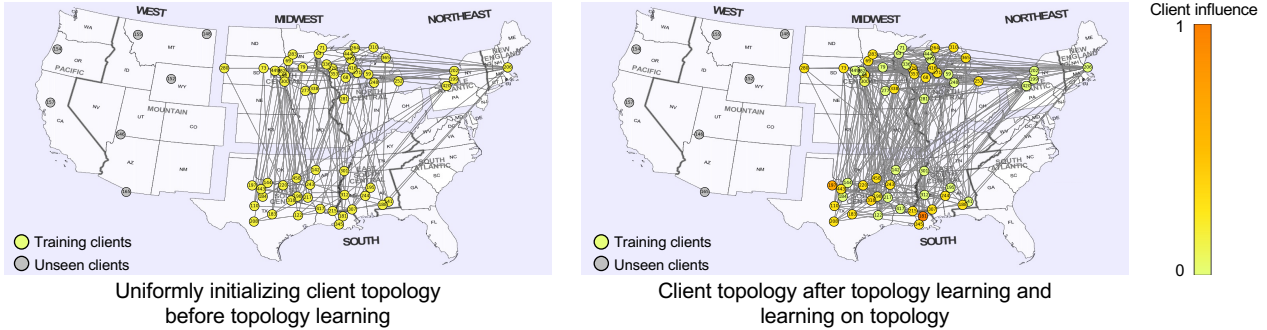
Figure 6. Visualization of client topology on the eICU dataset. Hospitals from MIDWEST, SOUTH, and NORTHEAST collaboratively train a global model, which is then applied to hospitals in the WEST region for OOF generalization. We observe that the learned client topology becomes denser, revealing underlying relationships that were previously unknown. Furthermore, the identified "influential" hospitals are more from the MIDWEST and SOUTH rather than NORTHEAST. This observation is rational, given the geographical proximity of these two regions to the target evaluation region, the WEST.
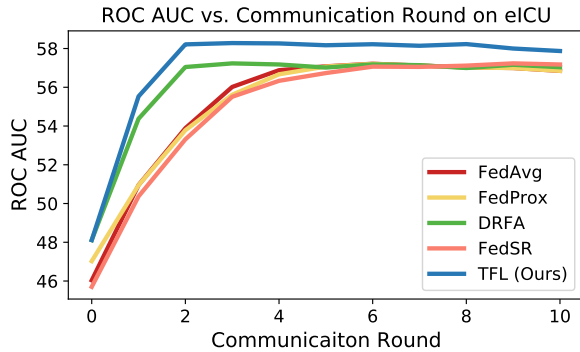

Figure 7. Visualization of ROC AUC vs. comm. round. *Our method yields the best result with fewer communication rounds.*

toward the optimum. Additionally, we also show the wall clock time versus OFF accuracy on PACS in Figure 2. Our method approximates the scalability of FedAvg and FedSR while achieving superior OOF accuracy.

4.4. Evaluation on In-federation Performance

Takeaway 3: Prioritizing influential clients also benefit in-federation (IF) performance. We have shown that TFL can significantly boost the model's OOF robustness. This naturally leads us to examine its IF performance. We are interested in understanding if TFL can also improve the IF performance. To this end, we conducted empirical evaluations using the TPT-48 and CIFAR-10/-100. The TPT-48 dataset, offering additional hold-out data from IF clients, is particularly well-suited for evaluating IF performance. CIFAR-10 and -100 are widely recognized benchmarks for IF evaluation. The results, shown in Tables 4 and 5, reveal that our method indeed enhances the IF performance. For example, TFL outperforms DRFA by 2.2% on TPT-48 and 2.5% on CIFAR-10. These findings underscore the value of focusing on influential clients when building robust models with good IF and OOF performance.

Table 4. Results on TPT-48. We report results for both IF and OOF evaluations. The numbers are the mean of three independent runs. *Our method improves both the OOF and IF performance.*

Method	Centralized	FedAvg	FedProx	DRFA	TFL (Ours)
Out-of-federation Evaluation (MSE ↓)					
E(24) → W(24)	0.3998	0.6264	0.6312	0.5451	**0.4978**
N(24) → S(24)	1.4489	2.0172	1.9729	1.8972	**1.7432**
In-federation Evaluation (MSE ↓)					
E(24) → E(24)	0.1034	0.2278	0.2163	0.1554	**0.1523**
N(24) → N(24)	0.1329	0.1550	0.1523	0.1437	**0.1405**

Table 5. In-federation accuracy on CIFAR-10/-100, where a Dirichlet distribution (Hsu et al., 2019) with parameter Dir(0.1) is used to partition the datasets into heterogeneous clients at different scales. The reported numbers are the mean of three independent runs. *Our method yields the best accuracy at various scales.*

Dataset	# of clients	FedAvg	FedProx	DRFA	TFL (Ours)
CIFAR-10	25	68.07	68.15	70.51	**72.24**
	50	64.48	64.33	63.19	**65.59**
CIFAR-100	25	37.90	37.76	38.04	**38.85**
	50	37.33	37.28	37.54	**38.08**

4.5. Evaluation on Effectiveness of Client Clustering

Takeaway 4: Client clustering reduces the computation costs of client topology learning. We conducted experiments on eICU to validate whether client clustering effectively reduces computation costs. The eICU dataset was chosen due to its larger scale (including 72 clients) compared to all other evaluated datasets. Specifically, during client topology learning, we employed KMeans (Lloyd, 1982) to group training clients into several clusters (*e.g.,* 10 clusters) based on their model weights. The client topology is then learned at the cluster level. As shown in Table 6, our clustering approach significantly reduces computation costs while slightly decreasing the OOF accuracy.

Table 6. Comparison of wall-clock time on eICU. *Our client clustering approach (with 10 clusters) significantly decreases computation costs by* 69%, *while only marginally impacting OOF performance, evidenced by a slight reduction of* 0.77%.

	ROC-AUC	Wall-clock time (s)
FedAvg	57.18±0.03	120.15
TFL	58.41±0.06	437.61
TFL w/ Clustering	57.96±0.18	133.08

4.6. Ablation Study

Effects of partial model approximation. To handle high-dimensional models, we can leverage partial model parameters to compute the similarity scores for better computation efficiency. Here, we report the OOF performance when using full and partial model parameters. Experiments are conducted on PACS and results are shown in Table 7.

Effects of graph sparsity. To avoid learning a fully connected graph with trivial edges, we add a sparsity contain to client topology learning. In implementation, we adopt ϵ-graph to ensure Equation 4 is solvable. The threshold value ϵ controls the graph sparsity. Here, we investigate how it will affect the TFL. We report the results on PACS in Table 8. We observe that $\epsilon = 0.4$ yields the best performance.

Effects of topology update frequency. The frequency of updating client topology during training affects the total training time ($f = 5$ denotes updating the topology every 5 rounds). More frequent updates result in extended training times. We investigated the impact of client topology updating frequency on the performance of the model on PACS. Our results in Table 8 indicate that model performance remains relatively stable across various updating frequencies. **We provide more ablation studies in Supplementary C.**

5. Related Work

FL generalization to unseen clients. There are recent attempts to address generalization to unseen clients in FL. FedDG (Liu et al., 2021) is proposed to share the amplitude spectrum of images among local clients to augment the local data distributions. FedADG (Zhang et al., 2021) adopts the federated adversarial training to measure and align the local client distributions to a reference distribution. FedSR (Nguyen et al., 2022) proposes regularizing latent representation's ℓ_2 norm and class conditional information to enhance the OOF performance. However, existing methods often ignore scalability issues, yielding inferior performance in large-scale distributed setting (Bai et al., 2023). We introduce an approach that employs client topology to achieve good OOF-resiliency in a scalable manner.

FL with Graphs. One line of research is to combine FL with GNNs, leading to federated GNNs (Xie et al., 2021;

Table 7. TFL's OOF accuracy on PACS using full and partial model parameters. The results show that *using partial model parameters achieves performance on par with full parameters.*

	A	C	P	S	Avg.
Full	59.05±0.69	64.46±0.21	79.35±0.61	76.93±0.39	69.95
Partial	58.96±0.85	64.57±0.98	78.94±0.65	76.61±0.13	69.77

Table 8. Ablation study on effects of graph sparsity and client topology update frequency on PACS.

Graph Sparsity (ϵ), Accuracy ↑				
$\epsilon = 0.45$	$\epsilon = 0.4$	$\epsilon = 0.38$	$\epsilon = 0.35$	$\epsilon = 0.3$
58.61	**59.05**	58.83	58.11	57.73
Topology update frequency (f), Accuracy ↑				
$f = 50$	$f = 30$	$f = 20$	$f = 10$	$f = 5$
59.19	58.91	58.27	58.80	58.28

Wu et al., 2021). These federated GNNs primarily engage in learning from distributed graph data that already has inherent structural information. In contrast, our primary objective is to obtain a distribution graph that captures the relationships between distributed clients. Another research line explores the use of graphs to enhance personalization in FL (Ye et al., 2023; Zantedeschi et al., 2019). For example, pFedGraph (Ye et al., 2023) utilizes a collaboration graph to tailor the global model more effectively to individual clients. These types of research primarily deal with in-federation clients. Different from these methods, our approach aims to utilize a client topology to improve the FL model's generalizability to clients outside of the federation.

Graph topology learning. The problem of graph topology learning has been studied in graph signal processing (Stanković et al., 2020) and graph neural networks (Chen et al., 2020a;b). Traditional topology learning methods often require centralizing the data, raising privacy concerns. Yet, how to estimate the graph topology with privacy regulations has been less investigated. We explore simple methods to infer the graph topology using model weights. **An extended related work section is provided in Supplementary B.**

6. Conclusion

Federated Learning faces significant out-of-federation (OOF) generalization challenges that can severely impair model performance. We propose to improve OOF robustness by leveraging client relationships, leading to *Topology-aware Federated Learning* (TFL). TFL contains two key modules: i) Inferring a topology that describes client relationships with model similarity and ii) Leveraging the learned topology to build a robust model. Extensive experiments on real-world and benchmark datasets show that TFL demonstrates superior OOF-resiliency with scalability.

Acknowledgment

Impact Statement

Societal impact. Our research has the potential to impact various sectors, notably automated vehicles (AVs), healthcare, and finance, through the enhancement of Federated Learning (FL) models. By improving the robustness of FL models for unseen clients, we anticipate a range of societal benefits. In healthcare, this could manifest as more accurate medical diagnostics, leading to better patient outcomes. In the realm of AVs, enhanced FL models could result in safer navigation and operation, particularly under challenging weather or road conditions. In finance, these advancements could foster better inter-company collaboration, leading to more efficient and secure financial services.

Ethical considerations. While we initially did not identify the potential for our research to lead to unintended biases or discrimination, we recognize the complexity and evolving nature of ethical challenges in Machine Learning. It is important to continuously evaluate our models and methodologies for any inadvertent biases, especially as they are applied in diverse, real-world scenarios. We commit to rigorously testing our models to identify and mitigate any such biases, ensuring that our research contributes positively and equitably across all sections of society.

References

Bai, R., Bagchi, S., and Inouye, D. I. Benchmarking algorithms for domain generalization in federated learning, 2023. URL https://openreview.net/forum?id=IsCg7qoy8i9.

Bavelas, A. Communication patterns in task-oriented groups. *The journal of the acoustical society of America*, 22(6):725–730, 1950.

Bhowmick, A., Duchi, J., Freudiger, J., Kapoor, G., and Rogers, R. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018. URL https://arxiv.org/abs/1812.00984.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Chen, Y., Wu, L., and Zaki, M. J. Graphflow: Exploiting conversation flow with graph neural networks for conversational machine comprehension. *arXiv preprint arXiv:1908.00059*, 2019.

Chen, Y., Wu, L., and Zaki, M. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. *Advances in neural information processing systems*, 33:19314–19326, 2020a.

Chen, Y., Wu, L., and Zaki, M. J. Reinforcement learning based graph-to-sequence model for natural question generation. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=HygnDhEtvr.

Deng, Y., Kamani, M. M., and Mahdavi, M. Distributionally robust federated averaging. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15111–15122, 2020. URL https://arxiv.org/abs/2102.12660.

Dong, X., Thanou, D., Rabbat, M., and Frossard, P. Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*, 36(3):44–63, 2019. URL https://ieeexplore.ieee.org/document/8700665.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Franceschi, L., Niepert, M., Pontil, M., and He, X. Learning discrete structures for graph neural networks. In *International conference on machine learning*, pp. 1972–1982. PMLR, 2019.

Freeman, L. C. et al. Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology. Londres: Routledge*, 1:238–263, 2002.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pp. 2029–2037. PMLR, 2018.

Huang, L., Shea, A. L., Qian, H., Masurkar, A., Deng, H., and Liu, D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics*, 99:103291, 2019.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. URL https://www.nowpublishers.com/article/Details/MAL-083.

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.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Li, R., Wang, S., Zhu, F., and Huang, J. Adaptive graph convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11691.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a. URL https://ieeexplore.ieee.org/abstract/document/9084352.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020b. URL https://arxiv.org/abs/1812.06127.

Li, T., Qiao, F., Ma, M., and Peng, X. Are data-driven explanations robust against out-of-distribution data? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3821–3831, 2023.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020c. URL https://openreview.net/forum?id=HJxNAnVtDS.

Liu, Q., Chen, C., Qin, J., Dou, Q., and Heng, P.-A. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1013–1023, 2021.

Lloyd, S. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Lü, L. and Zhou, T. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., and Peng, X. SMIL: Multimodal Learning with Severely Missing Modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021.

Ma, M., Ren, J., Zhao, L., Testuggine, D., and Peng, X. Are Multimodal Transformers Robust to Missing Modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022.

Mateos, G., Segarra, S., Marques, A. G., and Ribeiro, A. Connecting the dots: Identifying network structure via graph signal processing. *IEEE Signal Processing Magazine*, 36(3):16–43, 2019. URL https://ieeexplore.ieee.org/document/8700659?arnumber=8700659.

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. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 691–706. IEEE, 2019. URL https://ieeexplore.ieee.org/document/8835269.

Mendieta, M., Yang, T., Wang, P., Lee, M., Ding, Z., and Chen, C. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8397–8406, 2022.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625. PMLR, 2019. URL https://proceedings.mlr.press/v97/mohri19a.html.

Newman, M. *Networks*. Oxford university press, 2018.

Newman, M. E. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.

Nguyen, A. T., Tran, T., Gal, Y., and Baydin, A. G. Domain invariant representation learning with domain density transformations. *Advances in Neural Information Processing Systems*, 34:5264–5275, 2021.

Nguyen, A. T., Torr, P., and Lim, S.-N. Fedsr: A simple and effective domain generalization method for federated learning. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/pdf?id=mrt90D00aQX.

Norcliffe-Brown, W., Vafeias, S., and Parisot, S. Learning conditioned graph structures for interpretable visual question answering. *Advances in neural information processing systems*, 31, 2018.

Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.-H., Reina, G. A., Foley, P., Gruzdev, A., Karkada, D., Davatzikos, C., et al. Federated learning enables big data for rare cancer boundary detection. *Nature communications*, 13(1):7346, 2022a.

Pati, S., Baid, U., Edwards, B., Sheller, M. J., Foley, P., Reina, G. A., Thakur, S., Sako, C., Bilello, M., Davatzikos, C., et al. The federated tumor segmentation (fets) tool: an open-source solution to further solid tumor research. *Physics in Medicine & Biology*, 67(20):204002, 2022b.

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Qiao, F. and Peng, X. Topology-aware robust optimization for out-of-distribution generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Reisizadeh, A., Farnia, F., Pedarsani, R., and Jadbabaie, A. Robust federated learning: The case of affine distribution shifts. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21554–21565, 2020. URL https://arxiv.org/abs/2006.08907.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020. URL https://www.nature.com/articles/s41746-020-00323-1.

Robey, A., Pappas, G. J., and Hassani, H. Model-based domain generalization. *Advances in Neural Information Processing Systems*, 34:20210–20229, 2021. URL https://openreview.net/forum?id=JOxB9h40A-1.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Sattler, F., Wiedemann, S., Müller, K.-R., and Samek, W. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019. URL https://ieeexplore.ieee.org/document/8889996.

Sattler, F., Müller, K.-R., and Samek, W. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020. URL https://ieeexplore.ieee.org/document/9174890.

Saxena, A. and Iyengar, S. Centrality measures in complex networks: A survey. *arXiv preprint arXiv:2011.07190*, 2020.

Shamsian, A., Navon, A., Fetaya, E., and Chechik, G. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, pp. 9489–9502. PMLR, 2021. URL http://proceedings.mlr.press/v139/shamsian21a/shamsian21a.pdf.

Sheikhalishahi, S., Balaraman, V., and Osmani, V. Benchmarking machine learning models on multi-centre eicu critical care dataset. *Plos one*, 15(7):e0235424, 2020.

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):1–12, 2020.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

Staib, M. and Jegelka, S. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Stanković, L., Mandic, D., Daković, M., Brajović, M., Scalzo, B., Li, S., Constantinides, A. G., et al. Data analytics on graphs part iii: machine learning on graphs, from graph topology to applications. *Foundations and Trends® in Machine Learning*, 13(4):332–530, 2020.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016.

Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., and Zhang, C. FedProto: Federated Prototype Learning Across Heterogeneous Clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8432–8440, 2022.

Tenison, I., Sreeramadas, S. A., Mugunthan, V., Oyallon, E., Belilovsky, E., and Rish, I. Gradient masked averaging for federated learning. *arXiv preprint arXiv:2201.11986*, 2022.

Vanhaesebrouck, P., Bellet, A., and Tommasi, M. Decentralized collaborative learning of personalized models over networks. In *Artificial Intelligence and Statistics*, pp. 509–517. PMLR, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. URL https://arxiv.org/abs/1706.03762.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Vose, R., Applequist, S., Squires, M., Durre, I., Menne, M., Williams Jr, C., Fenimore, C., Gleason, K., and Arndt, D. Gridded 5km GHCN-Daily Temperature and Precipitation Dataset (nCLIMGRID) Version 1. *Maximum Temperature, Minimum Temperature, Average Temperature, and Precipitation*, 2014.

Welling, M. and Kipf, T. N. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*, 2016.

White, D. R. and Borgatti, S. P. Betweenness centrality measures for directed graphs. *Social networks*, 16(4): 335–346, 1994.

Wu, C., Wu, F., Cao, Y., Huang, Y., and Xie, X. Fedgnn: Federated graph neural network for privacy-preserving recommendation. *arXiv preprint arXiv:2102.04925*, 2021.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

Xie, H., Ma, J., Xiong, L., and Yang, C. Federated graph classification over non-iid graphs. *Advances in Neural Information Processing Systems*, 34:18839–18852, 2021.

Xu, M., Zhang, J., Ni, B., Li, T., Wang, C., Tian, Q., and Zhang, W. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6502–6509, 2020.

Xu, Z., He, H., Lee, G.-H., Wang, B., and Wang, H. Graph-relational domain adaptation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=kcwyXtt7yDJ.

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.

Yu, Y., Yang, Z., Wei, A., Ma, Y., and Steinhardt, J. Predicting out-of-distribution error with the projection norm. In *International Conference on Machine Learning*, pp. 25721–25746. PMLR, 2022.

Zantedeschi, V., Bellet, A., and Tommasi, M. Fully decentralized joint learning of personalized models and collaboration graphs. *arXiv preprint arXiv:1901.08460*, 2019.

Zhang, L., Lei, X., Shi, Y., Huang, H., and Chen, C. Federated learning with domain generalization. *arXiv preprint arXiv:2111.10487*, 2021.

Zhao, S., Gong, M., Liu, T., Fu, H., and Tao, D. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems*, 33:16096–16107, 2020.

Zhou, Z., Azam, S. S., Brinton, C., and Inouye, D. I. Efficient federated domain translation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=uhLAcrAZ9cJ.

Zhu, Y., Xu, W., Zhang, J., Du, Y., Zhang, J., Liu, Q., Yang, C., and Wu, S. A survey on graph structure learning: Progress and opportunities. *arXiv e-prints*, pp. arXiv–2103, 2021. URL https://arxiv.org/pdf/2103.03036.pdf.

Zhuang, W., Chen, C., and Lyu, L. When foundation model meets federated learning: Motivations, challenges, and future directions. *arXiv preprint arXiv:2306.15546*, 2023.

A. Additional Experimental Results & Detailed Dataset Description

A.1. Evaluating OOF-resiliency on OfficeHome

We conduct additional experiments on **OfficeHome** (Venkateswara et al., 2017) dataset. It contains $15,588$ images from four domains: art, clipart, product, and real world. The task is a 65-class classification problem. Like PACS's experimental setup, we evenly split each domain into 5 subsets, yielding 20 subsets, and treat each subset as a client. We followed the common *leave-one-domain-out* experiment, where 3 domains are used (15 clients) for training and 1 domain (5 clients) for testing. We use ResNet50 (He et al., 2016) as our model and train the model for 100 communication rounds. Each local client optimized the model using stochastic gradient descent (SGD) with a learning rate of 0.01, a momentum of 0.9, weight decay of $5e^{-4}$, and a batch size of 64. The model is evaluated using classification accuracy.

Table 9. Accuracy on the **OfficeHome** dataset. We conduct experiments using a *leave-one-domain-out* approach, meaning each domain serves as the evaluation domain in turn. Existing methods typically consider each domain as an individual client (Liu et al., 2021; Nguyen et al., 2022). However, in order to simulate a large-scale distributed setting, we took a different approach by further dividing each domain into 5 subsets and treating each subset as a separate client. This increased the total number of clients to 20. *Our method outperformed others across all experimental settings*, demonstrating superior results.

Models		Backbone	OfficeHome				
			A	C	P	R	Average
Centralized	Mixup (Xu et al., 2020)	ResNet50	64.7	54.7	77.3	79.2	69.0
Methods	CORAL (Sun & Saenko, 2016)	ResNet50	64.4	55.3	76.7	77.9	68.6
	FedAvg	ResNet50	24.10	23.16	40.19	43.47	32.73
Federated	FedProx	ResNet50	23.16	23.47	41.08	42.66	32.59
Learning	DRFA	ResNet50	25.29	23.98	41.23	42.35	33.21
Methods	FedSR	ResNet50	23.51	22.93	39.30	41.48	31.81
	TFL (Ours)	ResNet50	**26.37**	**24.47**	**43.96**	**44.74**	**34.89**

A.2. Data Pre-processing on eICU

We follow (Huang et al., 2019) to predict patient mortality using drug features. These features pertain to the medications administered to patients during the initial 48 hours of their ICU stay. We've extracted pertinent patient and corresponding drug feature data from two primary sources: the 'medication.csv' and 'patient.csv' files. Our final dataset is a table with the dimension of 19000×1411. Each row in this matrix symbolizes a unique patient, while each column corresponds to a distinct medication.

A.3. More Results on eICU

We conduct more experiments on the eICU dataset to evaluate the gap between in-federation (IF) and out-of-federation (OOF) and visualize the results in Figure 8. *We observe that existing FL methods are not robust against OOF data.*

A.4. Detailed Dataset Description

① **eICU** (eICU Collaborative Research Database) (Pollard et al., 2018) is a large-scale multi-center critical care database. It contains high granularity critical care data for over $200,000$ patients admitted to 208 hospitals across the United States. Each hospital is considered an individual client. We follow (Huang et al., 2019) to prepare the data. The final dataset includes data from approximately 19000 patients over 72 hospitals. The generalization task could be deploying models trained on hospitals from the SOUTH region to those in the WEST. The evaluation metric for patient mortality prediction is ROC-AUC.

② **FeTS** (Federated Tumor Segmentation Dataset) (Pati et al., 2022b) is a multi-institutional medical imaging dataset. It comprises clinically acquired multi-institutional MRI scans of glioma. A subset of the original data was used, comprising 358 subjects from 21 distinct global institutions. The associated task is to identify and delineate brain tumor boundaries. Each institution is a client. Evaluation metric is Dice Similarity Coefficient (DSC ↑). See more visualizations in Figure 9.

③ **TPT-48** (Vose et al., 2014) contains the monthly average temperature for the 48 contiguous states in the US from 2008 to
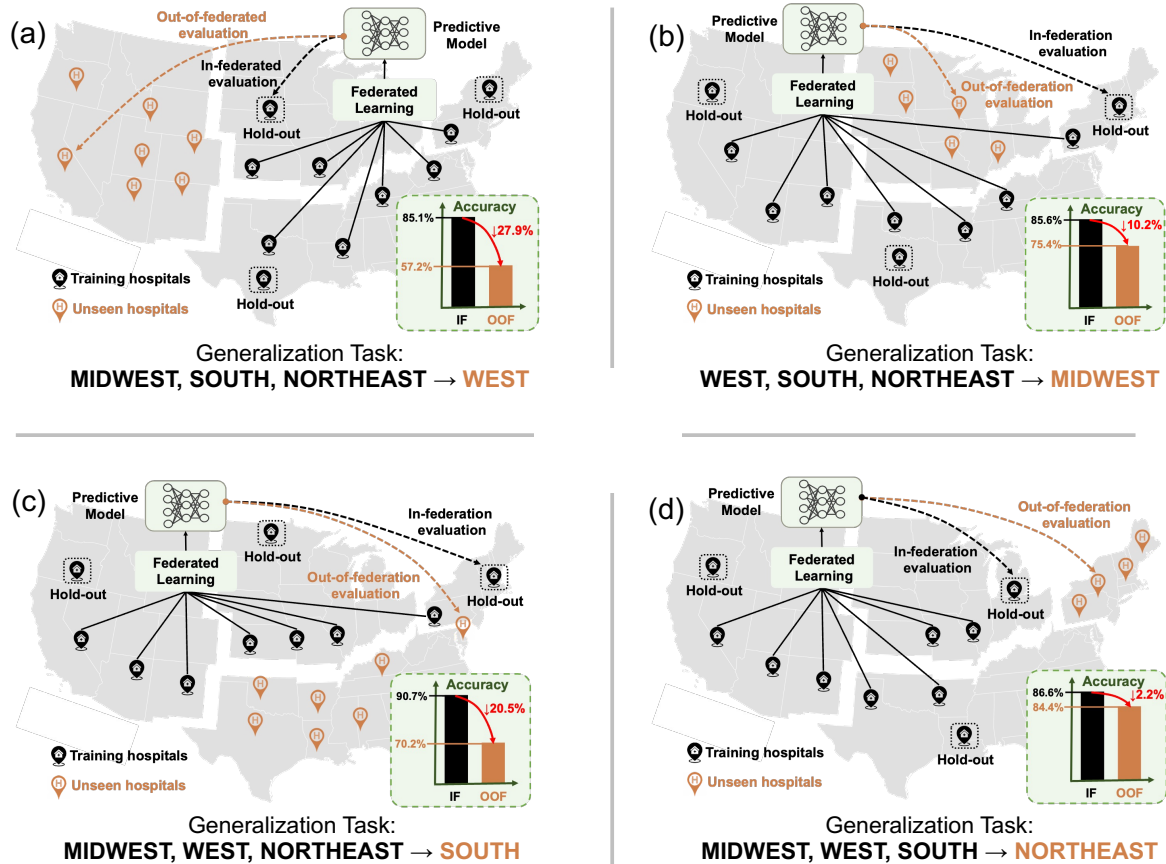
Figure 8. Additional results on eICU (IF *vs.* OOF performance). eICU's 72 hospitals are distributed across the United States. Specifically, there are 14 hospitals in the WEST, 28 hospitals in the MIDWEST, 26 hospitals in the SOUTH, and 4 hospitals in the NORTHEAST. We employ a *leave-one-region-out* approach, designating one geographic region as the OOF region while the remaining as IF regions. We observe a considerable gap between IF and OOF performance, indicating that *current FL methods are not robust against OOF data.*

2019. The task is to predict the next six months' temperature given the first six months' temperature. For TPT-48, we consider two generalization tasks: (1) E(24) → W(24): Using the 24 eastern states as IF clients and the 24 western states as OOF clients; (2) N(24) → S(24): Using the 24 northern states as IF clients and the 24 southern states as OOF clients. The evaluation metric is Mean Squared Error (MSE ↓).

④ **CIFAR-10/-100** (Krizhevsky & Hinton, 2009) are the most commonly used benchmarks in FL literature. In these datasets, we use Dirichlet distribution (Hsu et al., 2019) to partition the dataset into the heterogeneous setting with 25 and 50 clients.

⑤ **PACS** (Li et al., 2017) contains 9,991 images from four domains: art painting, cartoon, photo, and sketch. The task is seven-class classification. For PACS, we evenly split each domain into 5 subsets, yielding 20 subsets, and we treat each subset as a client. We followed the common "leave-one-domain-out" experiment, where 3 domains are used (15 clients) for training and 1 domain (5 clients) for testing. Model performance is evaluated by classification accuracy.

B. Extended Related Work

Federated learning. Federated learning (Li et al., 2020a; Kairouz et al., 2021; McMahan et al., 2017) has emerged as a powerful tool to protect data privacy in the distributed setting. It allows multiple clients/devices to collaborate in training a predictive model without sharing their local data. Despite the success, current FL methods are vulnerable to heterogeneous data (non-IID data) (Smith et al., 2017; Sattler et al., 2019; Li et al., 2023; Tan et al., 2022), a common issue in real-world FL. Data heterogeneity posits significant challenges to FL, such as the severe convergence issue (Li et al., 2020c) and poor generalization ability to new clients (Sattler et al., 2019). To improve the model's robustness against
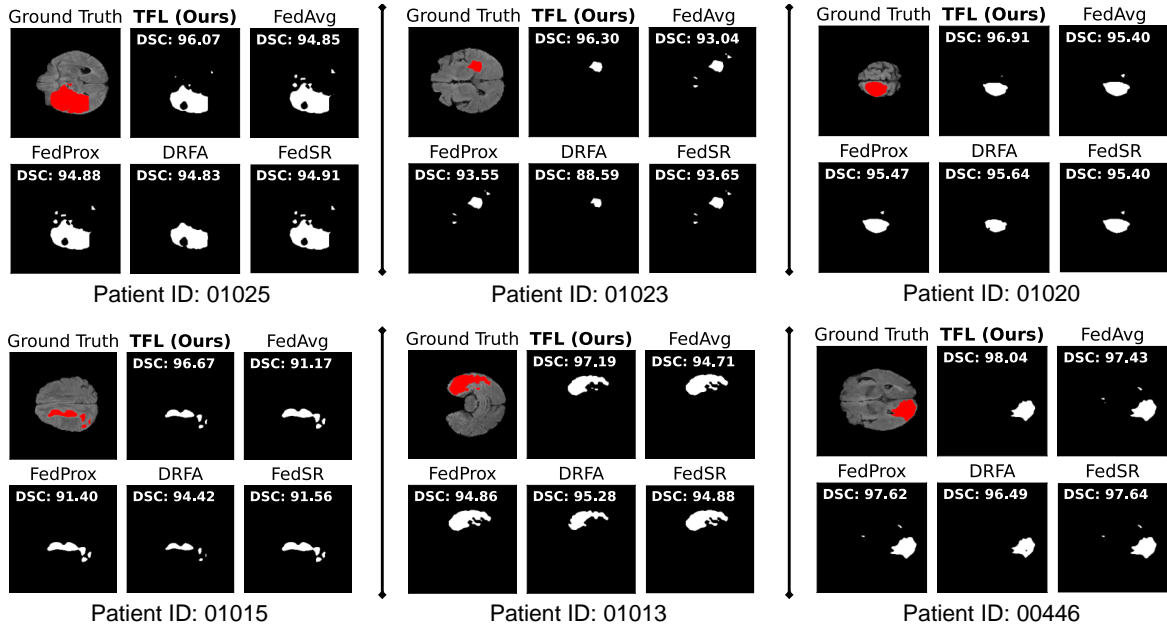
Figure 9. Additional qualitative results comparison on unseen patients of the FeTS dataset. We show both the tumor segmentation and DSC (↑) score. *Our method demonstrates consistent superior OOF-resiliency across a range of local demographics.*

data heterogeneity, `FedProx` (Li et al., 2020b) add a proximal term to restrict the local model updating, avoiding biased models toward local data distribution. `SCAFFOLD` (Karimireddy et al., 2020) introduces a control variate to rectify the local update. `FedAlign` (Mendieta et al., 2022) improve the heterogeneous robustness by training local models with better generalization ability. However, most FL methods focus on the model's in-federation performance. Orthogonal to existing work, we propose leveraging client topology to improve the model's OOF generalization capability.

FL generalization to unseen clients. There are recent attempts to address generalization to unseen clients in FL. FedDG (Liu et al., 2021) is proposed to solve domain generalization in medical image classification. The key idea is to share the amplitude spectrum of images among local clients to augment the local data distributions. FedADG (Zhang et al., 2021) adopts the federated adversarial training to measure and align the local client distributions to a reference distribution. FedGMA (Tenison et al., 2022) proposes gradient masking averaging to prioritize gradients aligned with the overall domain direction across clients. FedSR (Nguyen et al., 2022) proposes regularizing latent representation's ℓ_2 norm and class conditional information to enhance the OOF performance. However, existing methods often ignore scalability issues, yielding inferior performance in large-scale distributed setting (Bai et al., 2023). We introduce an approach that employs client topology to achieve good OOF-resiliency in a scalable manner.

Graph topology learning. The problem of graph topology learning has been studied in different fields. In graph signal processing (Mateos et al., 2019; Dong et al., 2019; Stanković et al., 2020), existing work explores various ways to learn the graph structure from data with structural regularization (*e.g.,* sparsity, smoothness, and community preservation (Zhu et al., 2021)). In Graph Neural Networks (GNNs) (Wu et al., 2020; Welling & Kipf, 2016), researchers have explored scenarios where the initial graph structure is unavailable, wherein a graph has to be estimated from objectives (Li et al., 2018; Norcliffe-Brown et al., 2018) or words (Chen et al., 2019; 2020b). The existing graph topology learning methods often require centralizing the data, making it inapplicable in federated learning. However, how to estimate the graph topology with privacy regulations has been less investigated. We explore simple methods to infer the graph topology using model weights.

FL with graphs. One line of research is to combine FL with GNNs, leading to federated GNNs (Xie et al., 2021; Wu et al., 2021). These federated GNNs primarily engage in learning from distributed graph data that already has inherent structural information. In contrast, our primary objective is to obtain a distribution graph that captures the relationships between distributed clients. Another research line explores the use of graphs to enhance personalization in FL (Ye et al., 2023; Zantedeschi et al., 2019). For example, pFedGraph (Ye et al., 2023) utilizes a collaboration graph to tailor the global model more effectively to individual clients. These types of research primarily deal with in-federation clients. Our approach, however, aims to utilize client topology to significantly improve generalization to clients outside of the federation.

C. Additional Ablation Study

Hyperparameter q**.** We investigate the impact of hyperparameter η on eICU. Our findings demonstrate that setting $q = 0.1$ yields the best results. **Centrality.** We employed betweenness centrality to derive the topological prior. However, it is worth noting that other types of centrality, such as degree (Freeman et al., 2002) and closeness (Bavelas, 1950), could also be utilized. We conducted experiments on eICU to verify the impact of different centrality measures on TFL. Our findings indicate that betweenness centrality produces the best result. **Similarity metrics.** We investigate how the model performs under different similarity metrics on PACS. We found that the dot product-based metric produces the best results.

Table 10. Ablation study evaluating the efficacy of hyperparameter tuning, centrality, and similarity metric.

Effectiveness of q, ROC AUC ↑				
$q = 1.0$	$q = 1e^{-1}$	$q = 1e^{-2}$	$q = 1e^{-3}$	$q = 1e^{-4}$
57.91	**58.31**	57.43	57.36	57.29
Effectiveness of centrality, ROC AUC ↑				
Betweenness	Degree	Closeness	Eigenvector	Current flow
58.28	57.69	57.86	57.57	57.83
Effectiveness of similarity measure, Accuracy ↑				
	ℓ_1	ℓ_2	dot produt	cosin
OOF Accuracy	58.11	58.26	59.14	58.52

D. Implementation Details

Experiment settings and evaluation metrics. For the **PACS dataset**, we evenly split each domain into 5 subsets, yielding 20 subsets, and we treat each subset as a client. We followed the common "leave-one-domain-out" experiment, where 3 domains are used (15 clients) for training and 1 domain (5 clients) for testing. We evaluated the model's performance using classification accuracy. We use ResNet18 (He et al., 2016) as our model and train the model for 100 communication rounds. Each local client optimized the model using stochastic gradient descent (SGD) with a learning rate of 0.01, momentum of 0.9, weight decay of $5e^{-4}$, and a batch size of 8. For **CIFAR-10/100**, we adopt the same model architecture as FedAvg (McMahan et al., 2017). The model has 2 convolution layers with 32, 64 5×5 kernels, and 2 fully connected layers with 512 hidden units. we use Dirichlet distribution (Hsu et al., 2019) to partition the dataset into the heterogeneous setting with 25 and 50 clients. For **eICU**, we treat each hospital as a client. We use a network of three fully connected layers. This architecture is similar to (Huang et al., 2019; Ma et al., 2021; Sheikhalishahi et al., 2020). We train our model for 30 communication rounds, using a batch size of 64 and a learning rate of 0.01, and report the performance on unseen hospitals. Within each communication round, clients perform 5 epochs (E = 5) of local optimization using SGD. The evaluation metric employed was the ROC-AUC, a common practice in eICU (Huang et al., 2019). For **FeTS**, we treat each institution as a client. We adopt the widely used U-Net (Ronneberger et al., 2015) model. We train our model for 20 communication rounds, using a learning rate of 0.01 and a batch size of 64. We conduct training with 16 intuitions and report results on 5 unseen institutions. Each institution performs 2 epochs of local optimization (E = 2) using SGD. The evaluation metric is Dice Similarity Coefficient (DSC ↑). For **TPT-48**, we consider two generalization tasks: (1) E(24) → W(24): Using the 24 eastern states as IF clients and the 24 western states as OOF clients; (2) N(24) → S(24): Using the 24 northern states as IF clients and the 24 southern states as OOF clients. We use a model similar to (Xu et al., 2022), which has 8 fully connected layers with 512 hidden units. We use SGD optimizer with a fixed momentum of 0.9. The evaluation metric is Mean Squared Error (MSE ↓). Algorithm 1 shows the overall algorithm of TFL. In implementation, we used dot product as the metric to measure client similarity. All experiments are conducted using a server with 8×NVIDIA A6000 GPUs.

E. Discussion of Limitations

In this section, we discuss the limitations of TFL and the potential solutions.

Concerns on privacy leakage. Client topology learning may raise concerns about (unintentional) privacy leakage. However, we argue that any such leakage would be a general issue for FL methods rather than a unique concern for our approach. In comparison to standard FL, our method does not require additional information to construct the client topology, thus providing no worse privacy guarantees than well-established methods like FedAvg (McMahan et al., 2017) and FedProx (Li et al., 2020b). Nonetheless, FL may still be vulnerable to attacks that aim to extract sensitive information (Bhowmick et al., 2018; Melis et al., 2019). In future work, we plan to explore methods for mitigating (unintentional) privacy leakage.

Concerns on high-dimensional node embedding. As outlined in Section 3.1, we harness model weights as node embeddings. Nevertheless, incorporating large-scale models, such as Transformers (Vaswani et al., 2017; Ma et al., 2022;

Algorithm 1 Topology-aware Federated Learning

Input: K clients; learning rate η_θ and η_λ; Communication round T; initial model $\theta^{(0)}$; initial $\lambda^{(0)}$.
repeat
  for each communication round $t = 1, \cdots T$ **do**
    server **samples** m clients according to $\boldsymbol{\lambda^{(t)}}$
    for each client $i = 1, \cdots m$ in parallel **do**
      $\theta_i^{t+1} = \theta_i^t - \eta_{\theta_i^t} \nabla_{\theta_i^t} F(\theta_i^t)$
      client i send θ_i^{t+1} back to the server
    end for
    server **computes** $\boldsymbol{\theta}^{t+1} = \sum_{i=1}^m \theta_i^{t+1}$
    if $t \% f == 0$ **then**
      server updating graph $\mathcal{G}$ via Equation 4
    end if
    server calculating topological prior p from $\mathcal{G}$
    server calculating $\nabla_{\lambda^{(t)}} F(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\lambda^{(t)}})$ via Equation 6
    server updates $\boldsymbol{\lambda}^{t+1} = \mathcal{P}_{\Delta_K}(\boldsymbol{\lambda}^t + \eta_\lambda^t \nabla_{\lambda^{(t)}} F(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\lambda^{(t)}}))$
  end for
until convergence

Dosovitskiy et al., 2021), may present a formidable obstacle, producing an overwhelmingly high-dimensional node vector. This will significantly increase computational demands for assessing node similarity. We argue that this can be addressed by dimension reduction. There are two possible ways: (1) Utilizing model weights of certain layers as node embedding instead of the whole model. (2) Directly learning the low-dimensional node embedding. One simple idea is to leverage Hypernetworks (Shamsian et al., 2021) to learn the node embedding with controllable dimensions.

F. Future Direction

Handling adversarial clients. We see the potential of using client topology to tackle adversarial clients. In scenarios where clients, due to system or network failures, disobey the protocols and send arbitrary messages (such as shuffled, sign-flipped, or noised parameters), our client topology learning approach becomes particularly useful. These adversarial behaviors typically result in models that show lower similarity to normal models. By leveraging client topology learning, we can identify these adversarial clients as isolated nodes within the topology. Once identified, applying centrality measures to these nodes can effectively lower their importance scores. This approach minimizes their impact on the overall model aggregation process. Therefore, we envision that our method holds the potential to tackle adversarial clients. Improving FL's adversarial robustness is an important and interesting problem, we will leave the exploration of TFL in this direction to future work.