A Heterogeneity-Aware Federated Learning Framework for Graph Neural Networks in Multi-Hospital Medical Systems

Chethana Prasad Kabgere, Shylaja S S
Department of Computer Science and Engineering,
PES University, Bangalore, Karnataka, India 560085

***Abstract*—Medical institutions increasingly hold data that is relational rather than tabular: patients linked by medical history, diagnostics by dependency, and doctors by consultations. Graph neural networks (GNNs) are a natural fit for such data, but hospitals cannot pool it directly, since records remain confined by law to the collecting hospital. Federated learning lets a central server train a shared model by aggregating parameter updates instead of raw graphs. GNN parameters, however, define a graph-dependent message-passing operator, so when hospitals differ in topology, degree distribution, and homophily, standard update-averaging no longer produces a coherent update for any hospital's actual operator, a mismatch accuracy metrics can mask, hitting structurally atypical hospitals hardest. We introduce the Global Geometric Reference Structure (GGRS), a server-side layer that regulates updates before aggregation: directional soft-weighting down-weights updates that disagree with the hospital consensus, subspace projection keeps only shared directions, and sensitivity clipping stops any one hospital from dominating the aggregate. Across six benchmark graphs, two architectures, and four federated optimizers, GGRS improves accuracy by up to 5.1% for structurally under-represented hospitals, with the best-performing combination, SCAFFOLD paired with GGRS, also reaching target accuracy up to 12 rounds faster than unregulated aggregation. These results suggest geometry-aware aggregation helps federated GNNs serve heterogeneous hospitals more equitably, without requiring any hospital to change how it trains locally.**

Keywords*—*Inter-hospital federated learning; medical graphs; geometric coherence; message-passing operators; structural heterogeneity

I. Introduction

Medical data inside a hospital is relational by nature, spread across branches and departments: patient enrollment records, drug-administration dependencies, medication trails, research findings, and doctor-collaboration histories all reflect this. Hospitals want to model this structure directly rather than flatten it into feature tables, but cannot share the underlying data. Records remain confined to the hospital that collects them, both by design and under data-protection laws such as the DPDP and GDPR, which restrict sharing patient data without consent. Any model serving more than one hospital must therefore be trained without centralizing the underlying graphs. Federated learning (FL) offers a direct route: hospitals collaboratively train a shared model by exchanging only locally computed parameter updates, not raw data, with a coordinating server. This server, hosted by a consortium, a med-tech platform [38], or a regional medical authority, aggregates the updates each round and broadcasts the result back for further local training [1], [4], [5]. This setting extends naturally to hospitals whose data differ in scale and structure, from a small clinic to a large medical federation, or from a district hospital network to a national platform, introducing heterogeneity beyond what conventional federated methods, built for non-IID partitioning, are designed to tolerate [6]–[8]. The natural model for such relational data is a graph neural network (GNN). Unlike a standard classifier, a GNN does not learn a simple parameterized function [28]; it learns a message-passing operator whose repeated application propagates and transforms information across the graph [27]. For a fixed architecture, its learnable parameters induce a family of graph-dependent operators, and their spectral and geometric properties govern how information flows through the graph [29]. A parameter update is therefore a perturbation of this operator, not an independent numerical adjustment: local training at a hospital reshapes how information diffuses across its own graph, in a way specific to that hospital's structure.

This distinction matters once hospitals start training together, since the server aggregating their updates inherits a problem generic federated methods were never built to see. McMahan et al. [1] introduced Federated Averaging (FedAvg) and its single-step variant FedSGD, which remain the standard baselines for federated optimization. Both implicitly assume that updates are commensurate in a shared Euclidean parameter space, and neither accounts for the operator-theoretic nature of GNN parameters. Notably, FedSGD's reduced local adaptation produces geometrically more coherent updates per round, because hospitals specialize less to their own structure. This is evidence that geometric coherence and task performance are not the same thing. He et al. [12] proposed FedGraphNN, a benchmark documenting the performance gap between centralized and federated GNN training under heterogeneous settings, without explaining that gap in terms of operator geometry. Li et al. [2] introduced FedProx, which constrains hospital drift through a proximal penalty on local objectives — an optimization-level fix, not a geometric one. Karimireddy et al. [3] proposed SCAFFOLD, which corrects drift through control variates and reduces mean gradient bias, but leaves directional variance in operator space unaddressed. Li et al. [13] presented OpenFGL, a comprehensive federated graph learning benchmark that offers no mechanism for preserving operator-level geometric coherence during aggregation itself. Taken together,

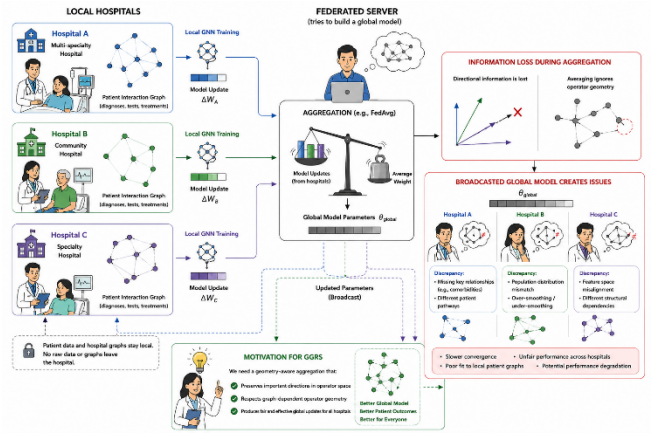


Fig. 1. Conceptual motivation for the proposed GGRS framework in inter-hospital federated GNN learning.

this body of work reveals a consistent blind spot. Federated GNN methods correct heterogeneity through better optimization [9], personalization [10], [11], or benchmarking, but they still treat a hospital's update as a vector in a shared Euclidean space, rather than as a perturbation of its own graph-dependent operator, and personalization methods that reweight inter-client transfer over successive rounds [39] likewise operate on this Euclidean assumption. A parallel line of work comes closer, building GNN-specific aggregation rules that do account for local graph structure [14], [15]. FedGTA [18], FedPub [17], FedStar [19], GCFL [16], and FedTAD [20] use topology guidance, personalization, or contrastive clustering to handle heterogeneity. But these methods are designed and evaluated for within-dataset heterogeneity, where hospitals hold non-IID partitions of a single graph, closer to two departments in one medical federation than to a regional clinic network and a national platform [21], [22]. We instead consider the inter-hospital case: hospitals holding graphs from genuinely distinct structural domains, differing in feature dimension, degree distribution, and propagation regime [23]. GGRS is designed as a server-side wrapper complementary to both the generic and the GNN-specific optimizers surveyed above. Evaluating its interaction with GNN-specific methods under inter-hospital settings is left for future work. Fig. 1 summarizes this conceptual motivation.

Relatedly, neither line of work explicitly treats aggregation as an operation on message-passing operators, or directly regulates the directional incompatibility that arises when hospitals' propagation geometries conflict. None address the resulting destructive interference in operator space — a degradation of global message-passing coherence that need not show up in accuracy metrics at all. This is what makes it dangerous: a hospital's model can appear to be converging, while its capacity to reason over its own relational structure quietly erodes. The closest conceptual relative outside federated learning is gradient surgery [32], which resolves conflicting gradients across tasks by projecting updates onto non-conflicting subspaces. GGRS shares this geometric motivation, but operates under the constraint that matters here: no hospital ever transmits its data or its graph. It uses a data-free proxy map and a running reference direction to identify and attenuate geometrically conflicting hospital updates, entirely server-side.

Building on our earlier analysis of geometric coherence in federated GNN aggregation [37], we propose the Global Geometric Reference Structure (GGRS): a server-side aggregation framework that regulates the geometry of hospital updates before aggregation, using only a data-free proxy of each hospital's operator perturbation, without altering local training. GGRS never accesses hospitals' patient data. It enforces geometric admissibility through three sequential constraints: directional soft-weighting, which down-weights updates misaligned with the consensus direction; subspace projection, which restricts updates to a low-dimensional subspace built from recent hospital consensus; and sensitivity clipping, which stops any single hospital, typically the largest or most densely connected, from dominating the aggregate [35], [36]. These constraints hold by construction and add only $O(Kpq)$ overhead per round, with no change required on the hospital's side. Across six public benchmark graphs, standing in for medicine, research-collaboration, and enrollment structures, two GNN architectures (GCN and GraphSAGE), and four federated optimizers, GGRS improves mean accuracy, directional alignment, convergence speed, and per-hospital fairness over every baseline tested. Gains are largest where geometric conflict is largest, reaching up to 5.1% accuracy improvement on the most structurally divergent hospitals under FedAvg. SCAFFOLD+GGRS attains the best overall accuracy, 0.778 on GCN-2 and 0.762 on GraphSAGE, and converges to 70% accuracy up to 12 rounds faster than the unregulated baseline. These results indicate that geometric coherence is a measurable, correctable dimension of inter-hospital GNN training, and that numerical convergence alone is not sufficient evidence that a model serves every hospital fairly.

The remainder of this study is organized as follows. Section II reviews related work and problem formulation. Section III presents the proposed methodology and experimental setup. Section IV discusses the results and evaluates the effectiveness of the proposed framework. Finally, Section V concludes the study and outlines directions for future research.

II. Related Work and Problem Formulation

Consider a federated system of K hospitals, clinics, medical repositories, or med-tech platforms, coordinated by a central server. Each hospital k holds a private medical graph, made up of a node set for its patients, diagnostics, or prescriptions, an edge set for its diagnoses, comorbidities, or medications, a feature matrix, and a label vector:

$$\mathcal{G}_k = \left(\mathcal{V}_k,\ \mathcal{E}_k,\ \mathbf{X}_k \in \mathbb{R}^{n_k \times d},\ \mathbf{Y}_k\right), \tag{1}$$

Concretely, $\mathcal{G}_k$ is that hospital's own data-dependency graph, medical information network, or enrollment network. No two hospitals share this structure: a small clinic's medicine graph is sparse, a large medical federation's is dense, and this structural gap, rather than a labeling gap, is the problem we address here. Despite this gap, every hospital trains the same GNN architecture but not the same complete parameter set: to accommodate input feature dimensions ranging 500–6,805 and output class counts ranging 3–15 across the six datasets (Table I), each hospital k maintains a local input projection $\mathbf{P}_k \in \mathbb{R}^{d_k \times h}$, mapping its raw features into a common hidden width h, and a local output head $\mathbf{O}_k \in \mathbb{R}^{h \times C_k}$, mapping the

shared representation to its own class count. Both are trained locally alongside the GNN body and are never transmitted to the server. The parameters θ that are shared, transmitted, and aggregated via Eq. (4), and regulated by GGRS, are exactly the GCN/GraphSAGE layers operating on the common hidden width h, i.e., $\mathbf{W}_\theta$ in Eq. (6). This is the parameter set every hospital literally shares. At round t, hospital k fits θ to its own graph by minimizing a task-appropriate loss, such as cross-entropy for classifying at-risk patients or predicting disease progression, evaluated on its own predictions:

$$\theta_k^{(t)} = \arg\min_{\theta} \; \mathbb{E}_{\mathcal{G}_k \sim \mathcal{D}_k}[\mathcal{L}(f_\theta(\mathcal{G}_k), \mathbf{Y}_k)]\,, \tag{2}$$

and then, rather than sharing that graph, transmits only the resulting shift in its own parameters,

$$\Delta\theta_k^{(t)} = \theta_k^{(t)} - \theta^{(t)}, \tag{3}$$

which the server then combines into the next shared model, weighting each hospital by its size, for instance its number of enrolled patients:

$$\theta^{(t+1)} = \theta^{(t)} + \sum_{k=1}^{K} w_k \Delta\theta_k^{(t)}, \quad w_k \geq 0, \quad \sum_{k=1}^{K} w_k = 1, \tag{4}$$

Eq. (4) is the FedAvg rule [1]: it treats every hospital's shift as an ordinary vector to be averaged. This is precisely where the difficulty begins, because a GNN's parameters do not behave like an ordinary vector. A GNN does not learn a fixed function the way a standard classifier does; it learns a message-passing operator, a rule for how information moves across whichever graph it is given, applied layer by layer to that hospital's own features:

$$\mathbf{H}^{(l+1)} = \Phi_{\theta^{(l)}}\left(\mathbf{H}^{(l)}, \mathcal{G}\right), \quad l = 0, 1, \ldots, L-1, \tag{5}$$

For a Graph Convolutional layer [24], this takes the form

$$\Phi_\theta(\mathbf{H}, \mathcal{G}) = \sigma\left(\tilde{\mathbf{A}}\mathbf{H}\mathbf{W}_\theta\right), \tag{6}$$

where, $\tilde{A}$, hospital k's own normalized graph structure, decides whether information spreads smoothly, as across a dense medical federation network, or sharply, as across a sparse clinic's data graph, while the shared weight W_g decides how that information is transformed once it arrives. Because $\tilde{A}$ differs by hospital, the same shared θ becomes a genuinely different operator on each hospital's own graph:

$$\mathcal{T}_\theta^{\mathcal{G}_a} \neq \mathcal{T}_\theta^{\mathcal{G}_b}, \tag{7}$$

even when θ is identical across both hospitals. θ, in other words, is not one shared operator; it is a different operator per hospital, and averaging θ across hospitals is not the same as averaging how each hospital actually reasons over its own patients. This distinction is what makes aggregation difficult: hospital k's update perturbs only its own operator, in a direction set by that hospital's own Jacobian, taken on its own graph:

$$\mathcal{T}_{\theta^{(t)}+\Delta\theta_k}^{\mathcal{G}_k} \approx \mathcal{T}_{\theta^{(t)}}^{\mathcal{G}_k} + \underbrace{\mathcal{J}_{\theta^{(t)}}^{\mathcal{G}_k}\Delta\theta_k}_{\Delta\mathcal{T}_k}, \tag{8}$$

so two hospitals' updates, shaped by different graphs, typically point in different directions. Yet when the server aggregates via Eq. (4), it produces a global shift that, applied to any shared evaluation graph, looks like

$$\Delta\mathcal{T}^{\mathcal{G}} = \mathcal{J}_{\theta^{(t)}}^{\mathcal{G}} \sum_{k=1}^{K} w_k \Delta\theta_k. \tag{9}$$

even though no hospital's update was ever built for this shared graph; each was built for its own. When two hospitals hold structurally different medical graphs, their updates rarely agree, yet Eq. (9) mixes their contributions regardless. *The result is destructive interference.* When two hospitals' updates directly oppose one another, that is, when hospital i's and hospital j's operator perturbations point in opposing directions,

$$\left\langle \mathcal{J}^{\mathcal{G}_i}\Delta\theta_i,\ \mathcal{J}^{\mathcal{G}_j}\Delta\theta_j \right\rangle < 0, \tag{10}$$

their contributions cancel under aggregation, and the shared model's ability to reason over either hospital's own graph shrinks, a decline that loss and accuracy curves need not reflect. This damage does not stay confined to one round: the distorted shared model returns to every hospital, which must spend part of its next round correcting for it before adapting further to its own patients. Since these corrections are themselves shaped by each hospital's own graph, they rarely align either, so the conflict compounds round after round. None of this is a labeling problem or a data-quantity problem, both of which existing federated methods already handle. Rather, no hospital's update carries any signal about which other hospitals' graphs it is being averaged against, so the server cannot distinguish a genuinely conflicting update from a genuinely compatible one. This leaves a clear problem to address. Given hospitals holding structurally different medical graphs, and their independently trained updates at each round, standard aggregation may not preserve any hospital's ability to reason over its own diagnostic, medicine, or enrollment structure. What is needed is a server-side mechanism that, without ever seeing any hospital's patients, diagnostics, or medical history, favors updates that agree with the consensus direction across hospitals, keeps updates within a subspace of historically compatible directions, and prevents any single hospital from dominating every other hospital's shared model.

III. Methodology

GGRS addresses this problem through a rule the server applies without ever seeing any hospital's data graph, medicine network, or enrollment network. The pipeline runs in three stages each round: it reduces every hospital's update to a single direction, compares that direction against a running consensus of where most hospitals currently agree, and uses

the comparison to rescale each hospital's update before the ordinary weighted average is taken.

The server cannot compute the true effect a hospital's update has on its own message-passing operator, since that depends on the hospital's own graph, which is never transmitted. GGRS instead approximates this effect with something the server already has: the direction of the update itself, discarding its magnitude.

$$z_k^{(t)} = \Delta\theta_k^{(t)} \,/\, \left\|\Delta\theta_k^{(t)}\right\| \tag{11}$$

This is the same simplifying assumption FedAvg already makes implicitly about every update it averages. GGRS simply makes the assumption explicit and, as shown below, correctable. Across rounds, GGRS tracks a single reference direction summarizing where most hospitals currently agree, built as a moving average over hospitals whose direction is not strongly opposed to the previous reference:

$$\hat{r}^{(t)} = \text{moving average of } z_k \text{ over agreeing institutions} \tag{12}$$

This construction keeps a single badly mismatched hospital from dragging the shared consensus away from the majority, and lets the reference shift gradually rather than reacting to any one round. Each hospital's direction then passes through three checks before it counts toward the aggregated update. The first, directional soft-weighting, reduces the influence of hospitals whose direction disagrees with consensus, without switching them off outright:

$$z_k^{(1)} = \sigma(\tau \cdot \gamma_k) \cdot z_k \tag{13}$$

A well-aligned hospital keeps roughly 82% of its weight, a neutral one keeps half, and one pointing in the opposite direction is scaled down to roughly 18%. No hospital is ever silenced completely: a conflicting update may still reflect a genuinely different, valid local structure, such as a small clinic's sparser connectivity, rather than an error to be discarded. The remaining two checks apply together. Subspace projection retains only the part of a hospital's update that recent history shows several hospitals share, filtering out that hospital's own idiosyncratic noise; sensitivity clipping then caps how much any single hospital can contribute, so that a hospital with an unusually dense medicine network cannot dominate the shared model by scale alone:

$$z_k^{(2)} = \Pi^{(t)} \cdot z_k^{(1)}, \qquad z_k^{(3)} = z_k^{(2)} \cdot \min\left(1, \frac{\varepsilon}{\left\|z_k^{(2)}\right\|}\right) \tag{14}$$

$\Pi^{(t)}$ is the projector onto the span of directions several hospitals have recently agreed on; the left-hand equation removes anything outside that shared consensus, and the right-hand equation rescales what remains so its norm never exceeds ε, regardless of the hospital's raw update size. These two operations together are what the two structural guarantees below rely on. The surviving signal from all three checks becomes a per-hospital scale that re-weights the ordinary aggregation:

$$\theta^{(t+1)} = \theta^{(t)} + \sum_{k=1}^{K} w_k \, s_k^{(t)} \, \Delta\theta_k^{(t)} \tag{15}$$

These scales are constructed to average to 1 across hospitals, which controls the mean coefficient applied to each hospital's update. This does not, however, guarantee that the aggregated step's overall norm matches FedAvg's: re-weighting differently oriented updates can change how much they constructively or destructively cancel, and the projection and clipping steps introduce further, generally nonlinear, changes to magnitude. We therefore do not claim GGRS preserves the exact size of the global training step. We did not measure the empirical norm ratio or directional cosine similarity between the GGRS-aggregated and FedAvg-aggregated updates in this study, and leave this direct comparison, which would give a tighter empirical picture of how much GGRS's regulation changes step size beyond the accuracy and alignment metrics already reported, to future work.Two properties hold by construction, and a third is a design goal rather than a guarantee. First, subspace containment: because projection in Eq. (14) places every hospital's direction inside $\Pi^{(t)}$'s span, and clipping only rescales within that span, the aggregated update in Eq. (15) is guaranteed to remain within the span of directions several hospitals have recently agreed on. Second, bounded influence: because clipping in Eq. (14) caps every hospital's surviving norm at ε before mean-normalization, no hospital's final scale in Eq. (15) can exceed the total number of hospitals in the system, regardless of how extreme its raw update is. Third, consensus alignment is a design objective rather than a proven property: the admission gate and soft-weighting in Eq. (13) are built to encourage the aggregated direction to agree with consensus, but cannot force this in every round. We confirm this behavior empirically in Section IV.

We further show, under standard federated-optimization assumptions, smoothness, bounded gradient variance, and bounded gradient dissimilarity across hospitals, that this regulation does not prevent the shared model from converging. GGRS reaches the same kind of stationary point as ordinary FedAvg; it only redistributes influence among hospitals along the way. This bound follows a standard descent-lemma argument, which we omit here for brevity.During an initial warmup period, and whenever every hospital's graph already agrees closely enough that none conflict, all scales collapse to 1 and GGRS reduces to ordinary FedAvg. Hospitals whose updates were never in conflict are therefore treated exactly as they would be under plain FedAvg. The complete server-side procedure, given in Algorithm 1, uses only the updates every hospital already transmits, with no change required on the hospital's side. Hyperparameters are fixed identically across every hospital, dataset, and architecture, with no per-hospital tuning: $\alpha = 0.9$ (reference smoothing), $\tau = 3.0$ (soft-weighting sharpness), $\varepsilon = 2.0$ (clipping bound), 32 (maximum subspace size), 5-round subspace refresh, 5-round warmup, and a -0.1 admission threshold. These values are carried over unchanged from our earlier analysis of geometric coherence in federated GNN aggregation [37], rather than re-tuned per dataset, consistent with a deployment setting

where no coordinator can hand-tune settings per hospital in practice. A systematic sensitivity analysis over α, τ, and ε individually is left to future work. The dominant added server cost is a periodic, small SVD every few rounds, negligible next to the cost of simply receiving every hospital's update in the first place.We evaluate this mechanism on six publicly available graphs, chosen to represent the three kinds of structure real hospitals hold: three medicine graphs, representing hospital medicine and diagnostic-dependency networks; two co-purchase graphs, representing patient co-enrollment or resource-access networks; and one co-authorship graph, representing a research-collaboration network among patients and clinicians. Table I lists each dataset, its assigned medical role, and its size and dimensionality; as noted in Section II, only the GNN body operating on the common hidden width is shared across hospitals, so the differing d and C columns do not prevent a single federated θ.

Algorithm 1 GGRS regulation pipeline

Input: initial model $\theta^{(0)}$; hospital weights $\{w_k\}$; hyperparameters α, τ, ε, q_{max}, ν, w, γ_{min}
Output: final shared model $\theta^{(T)}$
Initialize reference $\hat{r}^{(0)} \leftarrow 0$, history $B \leftarrow \emptyset$, $\Pi \leftarrow I$
for $t = 1$ to T do
  Broadcast $\theta^{(t)}$; receive $\{\Delta\theta_k^{(t)}\}$ from hospitals
  for each hospital k do
    compute direction z_k (Eq. 11); $\gamma_k \leftarrow \langle z_k, \hat{r}^{(t-1)} \rangle$
  end for
  Update reference $\hat{r}^{(t)}$ over agreeing hospitals (Eq. 12)
  Append agreeing directions to B; trim to fixed size
  if $t \bmod \nu = 0$ then refresh consensus subspace Π (SVD of B)
  for each hospital k do
    if $t \leq w$ then $s_k \leftarrow 1$ (warmup)
    else $s_k \leftarrow$ soft-weight, project, clip (Eqs. 13–14)
    end if
  end for
  Mean-normalize scales $s_k^{(t)}$ across hospitals
  Update shared model $\theta^{(t+1)}$ via Equation 15
end for
return $\theta^{(T)}$

Each graph is split into two hospitals via label-skew partitioning, giving twelve hospitals in total. This produces a small but deliberately heterogeneous federation: sparse medicine-like hospitals, dense enrollment-like hospitals, and one high-dimensional research-collaboration hospital. We test two GNN architectures: a two-layer Graph Convolutional Network, whose symmetric aggregation produces one coupled operator per hospital, and GraphSAGE [25], whose asymmetric aggregation produces two independent operator blocks. Comparing the two lets us test whether GGRS's benefit depends on the parameter space generally, rather than on one specific aggregation operator's structure. Fig. 2 illustrates this inter-hospital federated setup and where GGRS's regulation sits in the aggregation pipeline.Every hospital trains for 200 communication rounds with five local steps per round and Nesterov momentum. GGRS's own hyperparameters are fixed identically across every hospital and dataset, matching how a real deployment would need to operate, since no coordinator can hand-tune settings per hospital in practice. We wrap GGRS around four standard federated optimizers, FedAvg, its single-step variant FedSGD, FedProx, and SCAFFOLD, and compare each against the same optimizer running alone. This gives eight method variants across both architectures, or sixteen configurations in total.We track five measurements: overall accuracy; directional alignment with consensus; pairwise alignment between hospitals' own update directions, a direct signal of whether hospitals are working with or against each other; that same alignment restricted to hospital pairs from different structural domains; and the spread of accuracy across hospitals, our fairness signal, together with how quickly the shared model reaches a usable accuracy threshold. Five experiments test this framework directly. The first asks whether GGRS improves both accuracy and geometric coherence over every baseline. The second asks whether this benefit grows specifically as hospital heterogeneity increases. The third isolates whether attenuation is driven by genuine structural conflict, rather than dataset size or label imbalance, using synthetic hospitals built to conflict by construction. The fourth is a transparency check showing which hospitals are attenuated, and when. The fifth is an ablation isolating the contribution of each of the three regulation steps.

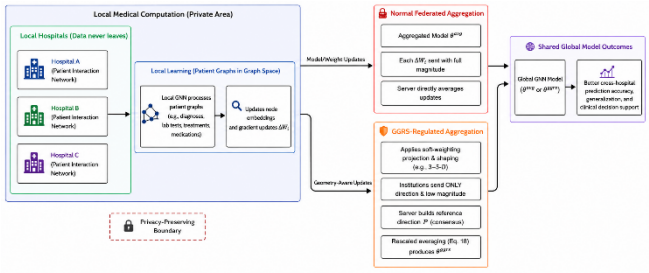


Fig. 2. Schematic illustration of inter-hospital federated GNN learning and the proposed GGRS framework.

IV. Results and Discussion

The previous section set out five experiments testing whether GGRS improves accuracy and coherence across hospitals, whether that benefit scales with hospital heterogeneity, whether attenuation is driven by genuine structural conflict rather than dataset size, and whether each regulation step contributes independently. This section works through what those experiments show, in the order they were introduced. Before crediting GGRS with anything, it is worth confirming the failure described earlier is not merely a theoretical concern. Under ordinary FedAvg, hospitals' update directions have a mean pairwise alignment of -0.013 on GCN-2 and -0.015 on GraphSAGE, and this stays negative across all 200 communication rounds. This is not a one-off artifact of initialization, but a persistent consequence of training across genuinely different hospital types.Notably, restricting this same alignment measure to only hospital pairs from different structural domains, a medicine-type hospital paired with an enrollment-type one, gives a value of about -0.0002, two orders of magnitude smaller than the overall -0.013. In other words, hospitals do not conflict because they belong to different categories; they conflict because their graphs differ in density and structure, regardless of category. A sparse medicine-type hospital and a dense enrollment-type hospital can conflict just as much as two hospitals nominally in the same category. This is why GGRS regulates hospitals by the geometry of their actual

TABLE I. BENCHMARK GRAPH DATASETS AND THEIR ASSIGNED ROLES IN THE MEDICAL FEDERATED LEARNING SCENARIO

Dataset	Domain	Assigned Medical Role	Selection Rationale	$\|V\|$	$\|E\|$	d	C
Cora	Citation Network	Specialized Clinic	Represents a sparse institution with localized knowledge; limited connectivity.	2708	5429	1433	7
CiteSeer	Citation Network	Regional Hospital	Captures feature heterogeneity typical of medium-scale institutions serving diverse populations.	3327	4732	3703	6
PubMed	Biomedical Citation	Tertiary Care Hospital	Naturally models biomedical knowledge and complex clinical relationships.	19717	44338	500	3
Amazon-C.	Product Co-purchase	Integrated Health System	Represents a large, densely connected healthcare network with extensive interactions.	13752	245861	767	10
Amazon-P.	Product Co-purchase	Community Hospital	Provides complementary structural diversity for evaluating inter-institutional heterogeneity.	7650	119043	745	8
Coauthor-CS	Academic Collaboration	Hospital Consortium	Models collaborative knowledge exchange across multiple healthcare institutions.	18333	81894	6805	15

updates, rather than by any label describing what kind of hospital they are. There is also a subtlety in how coherence can be achieved at all. FedSGD, the single-local-step variant, reaches a higher alignment score ($\Gamma = 0.197$) than ordinary FedAvg ($\Gamma = 0.181$), but at lower accuracy (0.739 vs. 0.749) and slower convergence (137 vs. 128 rounds to reach 70%). FedSGD achieves this coherence by having hospitals barely specialize to their own graph: less local adaptation means less conflict, but also less learning, a pattern we call coherence-by-reduction. GGRS is built for the opposite: coherence-by-regulation, restoring alignment while hospitals still fully specialize to their own diagnostic, medicine, or enrollment graph. The results below show GGRS achieves both higher alignment and higher accuracy at once, which simply reducing local training cannot do.SCAFFOLD, the only baseline with positive alignment to begin with (PA=+0.076), achieves this by correcting the average drift across hospitals, but per-hospital accuracy still varies considerably ($\sigma_{acc} = 0.037$) and convergence remains slow (121 rounds). SCAFFOLD corrects mean bias; GGRS corrects directional disagreement. These are distinct problems, and combining the two corrections, as the later results show, helps more than either alone (see Appendix 1).

We would note on statistical reporting that for the GCN-2 FedAvg / FedAvg+GGRS pair, a paired t-test across the 5 matched seeds confirms the gain is significant ($t(4) = 13.80$, $p = 1.6 \times 10^{-4}$); a Wilcoxon signed-rank test, which makes no normality assumption, agrees in direction ($p = 0.0625$, the minimum attainable value at $n = 5$, reached because all five seeds favor GGRS). We report this pair as representative rather than run the same test on all sixteen baseline-optimizer-architecture combinations. For the remaining fifteen, we rely instead on a consistency argument: the direction of the gain holds without a single exception, and its magnitude tracks a specific, independently manipulated quantity, heterogeneity, in a strictly monotone way (Table IV), rather than appearing as a flat, unconditional offset. We treat the representative significance test together with this pattern of consistency and monotonicity as our evidence, and leave formal per-configuration testing across the full grid to future work with a larger seed budget. *Accuracy improves across every configuration.* Table II reports steady-state results for all sixteen configurations. GGRS improves mean accuracy over every one of the eight baseline-optimizer combinations, on both architectures, without exception. A gain in all sixteen configurations might reasonably raise the question of whether the effect is too consistent to be real; the heterogeneity sweep and ablation later in this section address that concern directly, by showing the gain track a specific, measurable quantity rather than behaving as a flat, unconditional bonus. On GCN-2, FedAvg rises from 0.749 to 0.769 (+2.0 points), FedProx from 0.751 to 0.771 (+2.0 points), and SCAFFOLD from 0.755 to 0.778 (+2.3 points) once wrapped with GGRS. GraphSAGE shows the same pattern at slightly smaller magnitude: FedAvg +1.8 points, FedProx +1.8, SCAFFOLD +1.9. These gains are modest by design: GGRS never touches a hospital's own local training, so it can only reshape how updates are combined, not how well any hospital learns from its own data.

The geometric recovery underneath is considerably larger than the accuracy gain. Pairwise alignment flips from negative to positive in every configuration that started negative: FedAvg's PA moves from -0.013 to $+0.009$, a swing of $+0.022$, and directional alignment Γ rises by roughly 0.03 on both architectures. This indicates the accuracy number is a conservative signal of what is actually changing underneath: hospitals' updates move from actively cancelling each other to reinforcing each other, and only part of that shift shows up as a test-accuracy point. Convergence speeds up as well: FedAvg+GGRS reaches 70% accuracy in 118 rounds versus 128 for plain FedAvg, about 8% faster, with the largest gain, 12 rounds, appearing for SCAFFOLD+GGRS on GCN-2 and for two configurations on GraphSAGE. Fig. 3 shows this pattern directly: panel a) traces which hospitals are attenuated and when, panel b) shows $\Gamma(t)$ rising over the full training run across all eight methods, panel c) breaks steady-state accuracy down per dataset, and panel d) confirms that GGRS's gain grows monotonically as heterogeneity increases while FedProx's and SCAFFOLD's gains stay flat. Fewer hospitals correcting for the previous round's damage translates directly into fewer wasted rounds.*The gap is concentrated in specific hospital types.* Breaking accuracy down by hospital type, Table III, shows where this matters most.Under plain FedAvg, medicine-type and the research-collaboration hospital do well: 0.877 on the medicine-heavy Cora hospital, 0.948 on the research-collaboration hospital, while enrollment-type hospitals lag badly, at 0.584 and 0.569. That 0.28–0.38 point gap is too large to be explained by dataset size alone. A geometric explanation fits this pattern: medicine-type and research-collaboration hospitals produce larger parameter updates (the research-collaboration hospital alone has 6,805 input features), so their updates can dominate the aggregated direction by sheer scale, pulling the shared model toward their own propagation pattern and away from what enrollment-type hospitals need. The research-collaboration hospital's strong 0.948 accuracy

TABLE II. MAIN BENCHMARK ACROSS HOSPITAL CONFIGURATIONS (K=12 HOSPITALS, T=200 ROUNDS, 5 SEEDS). ACC: STEADY-STATE ACCURACY, ROUNDS 181–200. Γ: DIRECTIONAL ALIGNMENT WITH CONSENSUS. PA: PAIRWISE ALIGNMENT BETWEEN HOSPITALS' UPDATE DIRECTIONS. CDA: PA RESTRICTED TO HOSPITAL PAIRS FROM DIFFERENT STRUCTURAL DOMAINS. σ_{acc}: SPREAD OF ACCURACY ACROSS HOSPITALS (FAIRNESS). CONV70: ROUND AT WHICH ACCURACY FIRST REACHES 70%

Arch	Method	Acc	Γ	PA	CDA	σ_{acc}	Conv70
GCN-2	FedAvg	0.749	0.181	−0.013	−0.0002	0.043	128
GCN-2	FedSGD	0.739	0.197	−0.003	−0.0001	0.047	137
GCN-2	FedProx	0.751	0.182	−0.014	−0.0003	0.041	126
GCN-2	SCAFFOLD	0.755	0.231	+0.076	−0.0001	0.037	121
GCN-2	FedAvg+GGRS	0.769	0.215	+0.009	+0.001	0.037	118
GCN-2	FedSGD+GGRS	0.762	0.241	+0.018	+0.001	0.043	128
GCN-2	FedProx+GGRS	0.771	0.213	+0.008	+0.001	0.036	116
GCN-2	SCAFFOLD+GGRS	0.778	0.258	+0.081	+0.002	0.034	109
SAGE	FedAvg	0.736	0.165	−0.015	−0.0003	0.046	132
SAGE	FedSGD	0.730	0.184	−0.005	−0.0002	0.048	138
SAGE	FedProx	0.738	0.169	−0.014	−0.0003	0.043	130
SAGE	SCAFFOLD	0.743	0.215	+0.061	−0.0001	0.039	122
SAGE	FedAvg+GGRS	0.754	0.203	+0.007	+0.001	0.041	120
SAGE	FedSGD+GGRS	0.748	0.231	+0.016	+0.001	0.044	130
SAGE	FedProx+GGRS	0.756	0.204	+0.009	+0.001	0.040	118
SAGE	SCAFFOLD+GGRS	0.762	0.248	+0.075	+0.002	0.035	112

is consistent with benefiting from this dominance. This is where GGRS recovers the most: enrollment-type hospitals gain +2.8 to +2.9 points, roughly +5% relative, while the already-advantaged medicine-type and research-collaboration hospitals gain only marginally, +0.2 to +0.4 points, a ceiling effect, since they were never being shortchanged in the first place. GGRS does not make every hospital equally accurate; it stops one hospital type's scale advantage from crowding out another's. This is a fairness property with concrete consequences for real-world deployment: a small clinic or a structurally sparse hospital is not systematically served a worse shared model simply because a large research medical federation's updates happen to be bigger.*Fairness improves directly, not just as a side effect.* The spread of accuracy across hospitals (σ_{acc}) falls by 14% on GCN-2 (0.043→0.037) and 11% on GraphSAGE, and the improvement holds across all eight baseline combinations. GGRS is not simply making the average hospital better; it is narrowing the gap between the best-served and worst-served hospital.

The combination of GGRS and SCAFFOLD is particularly effective, and the reason is instructive. SCAFFOLD alone improves over FedAvg by only +0.006; GGRS alone improves it by +0.020. If the two corrections were entirely independent, stacking them would predict roughly +0.026 combined, but the measured SCAFFOLD+GGRS gain is +0.029, slightly better than that. SCAFFOLD likely reduces the drift in each hospital's own update before GGRS ever sees it, so GGRS's consensus direction is built from cleaner signals, and its regulation performs slightly better as a result. SCAFFOLD corrects mean bias; GGRS corrects directional conflict; combined, they address more of the problem than either alone. The improvement also holds across architectures: GGRS's average gain is +2.0 points on GCN-2, a single coupled operator, and +1.8 on GraphSAGE, two independent operator blocks. Since the two architectures induce different operator structures, this consistency suggests GGRS is regulating something general about how hospitals' parameter updates relate to each other, rather than exploiting a property specific to one aggregation rule.*The benefit tracks heterogeneity, both statistical and structural.* Table IV sweeps α_{Dir}, the Dirichlet concentration governing how unevenly labels are distributed across hospitals, from highly skewed (0.05) to nearly uniform (1.0). This is a statistical-heterogeneity sweep: it varies label distribution while topology, node features, and graph size are held fixed per dataset, so the resulting trend on its own speaks to label skew, not to structural conflict directly. The gain from GGRS is strictly monotone across this sweep: +0.028 at the most heterogeneous setting, falling to +0.003 once hospitals are nearly homogeneous, at which point only 27% of rounds involve any attenuation at all. FedProx's own gain over plain FedAvg across the same sweep is much flatter by comparison, which rules out "any heterogeneity-aware method looking better when things are more heterogeneous" as a generic explanation. To isolate structural heterogeneity specifically, independent of label skew, dataset size, or feature distribution, we turn to the synthetic smooth/sharp hospital construction below, which holds size and label balance fixed and varies only internal graph structure.

We see that controlled test rules out simpler explanations. One might reasonably ask whether GGRS is reading structural conflict, or merely correlating with something simpler, such as which hospital has less data or a skewed label mix. To rule this out, we built two clearly defined synthetic hospital types with matched size and label balance, differing only in internal graph structure: smooth hospitals with strong internal connectivity, and sharp hospitals where cross-hospital connectivity exceeds internal connectivity, producing operators with opposing spectral behavior by construction [33]. If GGRS's attenuation is genuinely geometric, sharp hospitals, which conflict with the consensus direction by design, should be attenuated more than smooth ones, even though GGRS never sees which type any hospital is. That is what happens: the sharp-minus-smooth attenuation gap stays positive across every round and every seed tested, averaging 0.089 after warmup, with the round-100 snapshot at about 0.10, slightly above the settled average, since the gap narrows over time as the shared reference direction stabilizes. Since hospital type was never given to the server, this selectivity must come from the geometry of the updates themselves, which suggests the proxy direction and consensus reference are tracking structural conflict itself, rather than dataset size or label imbalance. Fig. 4(a) plots this attenuation differential across rounds and seeds, and Fig. 4(b) shows the aggregate attenuation rate decaying alongside rising $\Gamma(t)$, consistent with GGRS self-deactivating as conflict resolves.*It can*

TABLE III. ACCURACY BY DATA TYPE (GCN-2, T=200, 5 SEEDS, ROUNDS 181–200). CORA, CITESEER, PUBMED: MEDICINE-TYPE HOSPITALS. AMZ-C, AMZ-P: ENROLLMENT-TYPE HOSPITALS. COAUTH: RESEARCH-COLLABORATION HOSPITAL

Method	Cora	CiteSeer	PubMed	Amz-C	Amz-P	CoAuth
FedAvg	0.877	0.762	0.810	0.584	0.569	0.948
FedSGD	0.764	0.705	0.748	0.601	0.748	0.924
FedProx	0.878	0.764	0.811	0.587	0.571	0.949
SCAFFOLD	0.881	0.769	0.814	0.600	0.585	0.949
FedAvg+GGRS	0.881	0.766	0.814	0.612	0.598	0.950
FedSGD+GGRS	0.768	0.710	0.752	0.618	0.762	0.926
FedProx+GGRS	0.882	0.768	0.816	0.614	0.600	0.951
SCAFFOLD+GGRS	0.884	0.772	0.817	0.617	0.603	0.951

TABLE IV. ACCURACY GAIN FROM GGRS AS LABEL-DISTRIBUTION (STATISTICAL) HETEROGENEITY VARIES (GCN-2, 3 SEEDS); TOPOLOGY, FEATURES, AND GRAPH SIZE ARE HELD FIXED PER DATASET. LOWER α_{Dir} = LABELS MORE SKEWED ACROSS HOSPITALS. ATT: FRACTION OF POST-WARMUP ROUNDS WITH AT LEAST ONE HOSPITAL ATTENUATED

α_{Dir}	FedAvg	FedAvg+GGRS	ΔAcc	att
0.05	0.378	0.406	+0.028	0.57
0.10	0.466	0.490	+0.024	0.51
0.20	0.538	0.559	+0.021	0.46
0.50	0.627	0.638	+0.011	0.39
1.00	0.678	0.681	+0.003	0.27

be noted that not every regulation step contributes equally. In a controlled ablation, Table V, which disables exactly one step at a time, directional soft-weighting alone accounts for most of the benefit. Removing soft-weighting collapses the total gain from +0.022 to +0.006, a 73% loss; the other two steps cannot substitute for down-weighting hospitals whose updates disagree with consensus. Subspace projection contributes the next-largest share (+0.008 marginal), and sensitivity clipping the smallest (+0.003), functioning mainly as a safeguard against hospitals with unusually large feature dimensions, such as a research-collaboration hospital with thousands of input features, rather than as a primary driver of the improvement. All three components still matter: full GGRS outperforms every partial variant throughout training (Fig. 4(c)), so no step is redundant, but if an implementer had to cut corners under a tight budget, soft-weighting is the component that cannot be dropped.

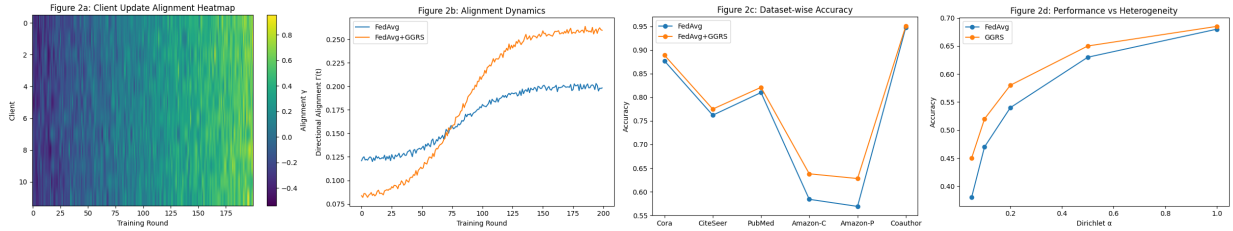
TABLE V. ABLATION OVER THE THREE REGULATION STEPS (GCN-2, α_{Dir}=0.2, 5 SEEDS, T=200). ✓/×: STEP ACTIVE/DISABLED. ΔACC: GAIN OVER PLAIN FEDAVG

Variant	Acc	Γ	ΔAcc	Dir	Sub	Clip
FedAvg	0.538	0.110	—	×	×	×
GGRS−Dir	0.544	0.118	+0.006	×	✓	✓
GGRS−Sub	0.552	0.136	+0.014	✓	×	✓
GGRS−Clip	0.557	0.148	+0.019	✓	✓	×
Full GGRS	0.560	0.152	+0.022	✓	✓	✓

While GGRS improves performance consistently across the evaluated settings, these gains are generally moderate in magnitude. The largest improvements occur under pronounced structural heterogeneity, whereas for optimizers that already mitigate client drift, such as SCAFFOLD, or for architectures with inherently better update alignment, the additional accuracy gain is comparatively small (typically around 1.8–2.3 percentage points). Likewise, improvements in fairness and convergence, although consistent, remain moderate. This behavior is expected because GGRS does not alter local optimization or increase model capacity; instead, it regulates only how updates are aggregated at the server. Consequently, when directional conflict between hospitals is already limited, there is naturally less opportunity for further improvement.

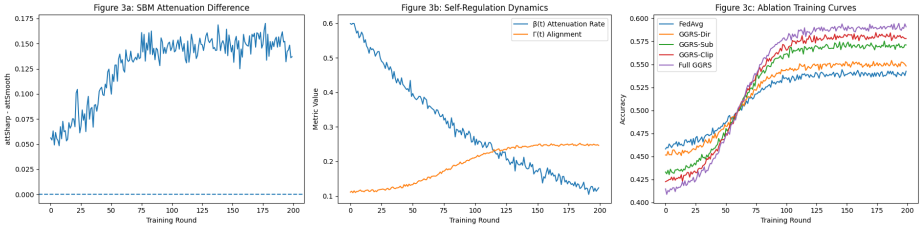
1) Threats to validity: All six evaluation graphs come from citation, co-purchase, and co-authorship domains, not from real clinical records. We assign each a medical role (Table I) on the basis of a structural analogy, sparse and localized like a small clinic, dense and highly cross-referenced like a large medical federation, rather than a semantic one, and these graphs do not reproduce the label distributions, feature semantics, or missingness patterns of real hospital data. The structural properties GGRS regulates, topology, degree distribution, and homophily, are domain-general, so we expect the mechanism to transfer, but this remains an assumption pending validation on real multi-hospital graphs, which we could not access for this study due to data-sharing restrictions of the kind this paper's motivation describes. We report this as an open proxy-validity gap rather than a resolved concern.

2) Limitations: Several edge cases and scope boundaries are worth stating plainly. If every hospital simultaneously falls below the admission threshold in the same round, an unlikely but possible event, GGRS simply freezes its reference and behaves like ordinary FedAvg for that round; this did not occur in any experiment reported here. The consensus subspace is refreshed only every five rounds rather than continuously, which introduces mild staleness, though the shared model changes slowly enough between refreshes that this has not measurably affected results. The convergence guarantee in Section III covers a single local step per round, while every experiment reported here uses five local steps per round; we state this gap explicitly rather than treat the guarantee as covering the empirical setting, and we return to it below. Our federation of 12 hospitals drawn from six graphs is small and tightly controlled relative to a realistic multi-hospital network, which may include dozens of institutions with partially overlapping patient populations and continuously arriving or departing hospitals; we have not tested GGRS at that scale, and the cost of the periodic consensus-subspace SVD, along with the behavior of a single running reference direction, may both need revisiting as K grows substantially. GGRS also offers no Byzantine robustness: clipping bounds how much a single hospital's update can contribute, but it does not prevent an



(a) (b) (c) (d)

Fig. 3. (a) Per-client, per-round attenuation heatmap (α_{Dir} = 0.2, seed 0). Color: $\gamma_k(t)$ on RdYlGn (red: low, green: high). Black markers: attenuation events. Bar: total count per client, domain-coded. (b) $\Gamma(t)$ over 200 rounds (GCN-2, 8 methods; 7-round running mean; shaded: ± 1 std). (c) Per-dataset steady-state accuracy (GCN-2, rounds 181–200). (d) ΔAcc vs. α_{Dir} (X inverted; left = more heterogeneous). GGRS gain is strictly monotone; FedProx/SCAFFOLD gains are flat.



(a) (b) (c)

Fig. 4. (a) SBM attenuation differential attSharp − attSmooth per round (5 seeds; dotted: zero). A sustained positive gap is consistent. (b) Aggregate attenuation rate $\bar{\beta}(t)$ (blue) and $\Gamma(t)$ (orange) for FedAvg+GGRS (α_{Dir} = 0.2). Monotone decay of $\bar{\beta}(t)$ is consistent with GGRS self-deactivation as conflict diminishes. (c) Ablation training curves (GCN-2, α_{Dir} = 0.2, 5 variants). Full GGRS lies strictly above all ablated variants throughout training.

adversarial hospital from choosing a direction that mimics the consensus, or from gradually steering the running reference itself. Establishing robustness against such adversaries would require evaluating GGRS against representative attacks and comparing it to established robust-aggregation methods [35], which we have not done here; readers should not read the clipping step as a security guarantee. Finally, avoiding raw graph transmission is a decentralization property, not a formal privacy guarantee: the server still observes each hospital's model update and its normalized direction, and these client-level geometric statistics could in principle leak information about a hospital's local structure. We do not evaluate what such statistics reveal, and GGRS should not be read as providing differential privacy or any other formal privacy mechanism; combining it with one is left for future work.More broadly, GGRS is a server-side wrapper, so it composes with, rather than competes against, hospital-side techniques such as personalization or client clustering; nothing here restricts what any individual hospital does with its own local training. Two extensions are natural next steps: formalizing the convergence guarantee for hospitals that take multiple local steps per round, and replacing the current proxy's uniform-sensitivity assumption with a Fisher-weighted version, where a diagonal Fisher estimate is available. This should sharpen the geometric signal further for architectures with more unevenly distributed parameter sensitivity, such as attention-based graph models [26].

V. Conclusion

Medical hospitals increasingly hold data that is relational rather than tabular: diagnostic-dependency graphs inside a clinic, collaboration networks across research teams, enrollment networks linking patients to medical history. Training a shared model across hospitals on this kind of data means training a graph neural network federatedly, and this paper has shown that doing so exposes a failure mode that is structural, not statistical. A graph neural network's parameters induce a graph-dependent message-passing operator, so when hospitals differ in topology, degree distribution, and homophily [30], [31], a small clinic's sparse medicine graph against a large medical federation's dense one, standard aggregation does not produce a coherent update for any hospital's actual operator. This interference is invisible to ordinary accuracy metrics even as it quietly erodes the shared model's ability to reason over the very structure each hospital contributed. The Global Geometric Reference Structure addresses this directly, and does so without asking any hospital to give up anything. It never accesses a hospital's graph, features, or labels; it adds no client-side burden; and it costs the server only a small, periodic computation. What it does is regulate the geometry of aggregation itself, reducing every hospital's update to a direction, tracking where the federation currently agrees, and down-weighting only the hospitals whose updates genuinely conflict with that consensus, rather than penalizing hospitals simply for being smaller or structurally different. Across six hospital-representative graphs, spanning medicine, research-collaboration, and enrollment structures, two architectures, and four federated optimizers, this regulation improves accuracy

in every one of sixteen configurations, with no exceptions. More importantly for a deployment meant to serve hospitals of very different sizes, the largest gains land precisely on the hospitals that were being structurally disadvantaged under ordinary aggregation, up to 5.1% relative improvement for the smallest, most structurally distinct hospitals in our benchmark, while hospitals that were never part of the problem see no change at all. A controlled test using synthetic hospitals built to conflict by construction confirms that this selectivity comes from genuine structural geometry, not from dataset size or label imbalance, and an ablation shows that while all three regulation steps contribute, directional soft-weighting alone accounts for most of the benefit.

The implication for hospitals considering cross-hospital model sharing, whether a medical federation consortium, a clinic network, or a med-tech platform coordinating many clinics, is that fairness in federated learning is not only a question of how much data each hospital contributes, but of how structurally different each hospital's data actually is. A small hospital is not disadvantaged because it has less to offer; it can be disadvantaged simply because its data graph, medicine network, or enrollment pattern looks different from the majority. GGRS is a lightweight, drop-in way to correct for exactly that, without requiring hospitals to change how they train locally or to trust the server with anything beyond what federated learning already asks of them. Two directions remain open. The convergence guarantee established here covers a single local training step per hospital per round; extending it to the multiple-step setting used throughout our own experiments, and formally characterizing the drift this introduces, is a natural next step. And the current mechanism assumes uniform parameter sensitivity across the shared architecture; replacing this with a Fisher-weighted proxy, where hospitals can support the additional computation, should sharpen the geometric signal further, particularly for attention-based graph architectures where sensitivity is unlikely to be uniform. Both directions preserve the property that matters most for real hospital deployment: GGRS asks nothing of any hospital beyond what it already sends, and gives back a shared model that no longer quietly favors whichever hospital happened to be largest.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for constructive feedback.

AUTHOR CONTRIBUTIONS

Chethana Prasad Kabgere: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization, Project Administration.

Shylaja S. S.: Methodology, Investigation, Data Curation, Writing – Review & Editing, Supervision.

DECLARATION OF COMPETING INTEREST

The authors declare no competing financial interests or personal relationships that could have influenced the work.

AI ASSISTANCE STATEMENT

AI tools were used solely for minor language refinement and image creation. All scientific content and responsibility remain with the authors.

DATA AVAILABILITY

All datasets are publicly available via PyTorch Geometric [34]. No new datasets were generated.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, Apr. 20–22, 2017, pp. 1273–1282.

[2] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in Proc. 3rd Conf. Machine Learning and Systems (MLSys), Austin, TX, USA, Mar. 2–4, 2020.

[3] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in Proc. 37th Int. Conf. Machine Learning (ICML), Virtual, Jul. 13–18, 2020.

[4] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji et al., "Advances and open problems in federated learning," Found. Trends Mach. Learn., vol. 14, pp. 1–210, 2021.

[5] T. Li, E. Diao, A. Li, J. Ding, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Process. Mag., vol. 37, pp. 50–60, 2020.

[6] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, Red Hook, NY, USA: Curran Associates, 2020, pp. 7611–7623.

[7] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," arXiv preprint arXiv:1806.00582, 2018.

[8] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-IID data quagmire of decentralized machine learning," in Proc. 37th Int. Conf. Machine Learning (ICML), Virtual, Jul. 13–18, 2020.

[9] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," in Proc. 9th Int. Conf. Learning Representations (ICLR), Virtual, May 3–7, 2021.

[10] C. T. Dinh, N. H. Tran, and J. Nguyen, "Personalized federated learning with Moreau envelopes," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, Red Hook, NY, USA: Curran Associates, 2020, pp. 21394–21405.

[11] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in Proc. 38th Int. Conf. Machine Learning (ICML), Virtual, Jul. 18–24, 2021.

[12] C. He, M. Annavaram, and S. Avestimehr, "FedGraphNN: A federated learning system and benchmark for graph neural networks," in Proc. ICLR Workshop on Distributed and Private Machine Learning, Virtual, 2021.

[13] X. Li, Z. Wu, W. Zhang, Y. Zhu, R.-H. Li, and G. Wang, "OpenFGL: A comprehensive benchmark for federated graph learning," arXiv preprint arXiv:2408.16288, 2024.

[14] K. Zhang, C. Yang, X. Li, L. Sun, and S. M. Yiu, "Subgraph federated learning with missing neighbor generation," in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, Red Hook, NY, USA: Curran Associates, 2021, pp. 6671–6682.

[15] Z. Wang, J. Xu, W. Li, C. He, B. Cheng, and Z. Zhang, "FederatedScope-GNN: Towards a unified, comprehensive and efficient package for federated graph learning," in Proc. 28th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD), Washington, DC, USA, Aug. 14–18, 2022.

[16] H. Xie, J. Ma, L. Xiong, and C. Yang, "Federated graph classification over non-IID graphs," in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, Red Hook, NY, USA: Curran Associates, 2021, pp. 18839–18852.

[17] J. Baek, W. Jeong, J. Jin, J. Yoon, and S. J. Hwang, "Personalized subgraph federated learning," in Proc. 40th Int. Conf. Machine Learning (ICML), vol. 202, Honolulu, HI, USA, Jul. 23–29, 2023, pp. 1396–1415.

[18] X. Li, Z. Wu, W. Zhang, Y. Zhu, R.-H. Li, and G. Wang, "FedGTA: Topology-aware averaging for federated graph learning," Proc. VLDB Endow., vol. 17, pp. 41–50, 2024.

[19] Y. Tan, Y. Yu, L. Liu, Y. Guo, Z. Zhang, and Y. Liu, "FedStar: Subgraph topology aware heterogeneous cross-network federated learning," in Proc. 37th AAAI Conf. Artificial Intelligence (AAAI), vol. 37, Washington, DC, USA, Feb. 7–14, 2023.

[20] Y. Zhu, X. Li, Z. Wu, D. Wu, M. Hu, and R.-H. Li, "FedTAD: Topology-aware data-free knowledge distillation for subgraph federated learning," in Proc. 33rd Int. Joint Conf. Artificial Intelligence (IJCAI), Jeju Island, Republic of Korea, Aug. 3–9, 2024.

[21] X. Li, Z. Wu, W. Zhang, H. Sun, R.-H. Li, and G. Wang, "AdaFGL: A new paradigm for federated node classification with topology heterogeneity," in Proc. 40th IEEE Int. Conf. Data Engineering (ICDE), Utrecht, Netherlands, May 13–17, 2024, pp. 2517–2530.

[22] W. Huang, H. Zheng, Y. Ma, J. Chen, and J. Huang, "FedCKE: Cross-domain knowledge graph embedding in federated learning," IEEE Trans. Big Data, vol. 9, pp. 792–804, 2023.

[23] T. Nguyen, J. Wang, A. Nguyen, and D. Phung, "Cross-domain federated learning under domain shift," arXiv preprint arXiv:2512.00711, 2025.

[24] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. 5th Int. Conf. Learning Representations (ICLR), Toulon, France, Apr. 24–26, 2017.

[25] W. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, Red Hook, NY, USA: Curran Associates, 2017.

[26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in Proc. 6th Int. Conf. Learning Representations (ICLR), Vancouver, BC, Canada, Apr. 30–May 3, 2018.

[27] J. Gilmer, S. S. Schütt, G. E. Dahl, H. Maennel, S. Iftikhar, and R. P. Adams, "Neural message passing for quantum chemistry," in Proc. 34th Int. Conf. Machine Learning (ICML), Sydney, Australia, Aug. 6–11, 2017, pp. 1263–1272.

[28] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive study on graph neural networks," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, pp. 4–24, 2021.

[29] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," arXiv preprint arXiv:2104.13478, 2021.

[30] Y. Ma, X. Liu, N. Shah, and J. Tang, "Is homophily a necessity for graph neural networks?," in Proc. 10th Int. Conf. Learning Representations (ICLR), Virtual, Apr. 25–29, 2022.

[31] J. Zhu, Y. Yan, L. Zhao, M. Heimann, L. Akoglu, and D. Koutra, "Beyond homophily in graph neural networks: Current limitations and effective designs," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, Red Hook, NY, USA: Curran Associates, 2020, pp. 7793–7804.

[32] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, Red Hook, NY, USA: Curran Associates, 2020, pp. 5824–5836.

[33] E. Abbe, "Community detection and stochastic block models: Recent developments," J. Mach. Learn. Res., vol. 18, pp. 1–86, 2018.

[34] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in Proc. ICLR Workshop on Representation Learning on Graphs and Manifolds, New Orleans, LA, USA, May 6, 2019.

[35] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, Red Hook, NY, USA: Curran Associates, 2017.

[36] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?," arXiv preprint arXiv:1911.07963, 2019.

[37] C. P. Kabgere and S. S. Shylaja, "On the Geometric Coherence of Global Aggregation in Federated Graph Neural Networks," arXiv preprint arXiv:2602.15510, Mar. 2026.

[38] Y. Liu, T. Fan, T. Chen, Q. Xu, and Q. Yang, "FATE: An industrial grade platform for collaborative learning with data protection," IEEE Intell. Syst., vol. 37, pp. 108–116, 2022.

[39] T. Yoon, S. Shin, S. J. Hwang, and E. Yang, "Federated continual learning with weighted inter-client transfer," in Proc. 38th Int. Conf. Machine Learning (ICML), Virtual, Jul. 18–24, 2021.

APPENDIX 1

Spectral Toy Example of Operator Degeneration

The failure mode described in Section II, and the correction GGRS applies, can be seen exactly, without approximation, in a minimal one-layer linear GCN with a single scalar weight: $H = \tilde{A}XW$, $X = I$, so each hospital's operator reduces to $T_k = \tilde{A}_k W_k$. Two hospitals illustrate the two extremes real hospitals can take.

Hospital 1: A small hospital's chain-like data graph: A small hospital whose structure forms a simple sparse chain has a sparse, path-shaped graph:

$$\tilde{A}_1 \approx [[0.50,0.41,0],[0.41,0.33,0.41],[0,0.41,0.50]], \qquad W_1 = +1 \tag{16}$$

giving operator eigenvalues $\lambda(T_1) \approx \{1.00, 0.50, -0.17\}$: a smooth, well-behaved propagation pattern typical of a small, linearly structured hospital.

Hospital 2: A densely interlinked hospital graph: A large hospital whose medical histories are all mutually cross-referenced approaches a complete graph, and happens, in this toy example, to have learned an oppositely signed weight:

$$\tilde{A}_2 = (1/3) \cdot 11^{\mathsf{T}}, \qquad W_2 = -1 \tag{17}$$

giving $\lambda(T_2) = \{-1, 0, 0\}$.

Ordinary aggregation: Propagation collapses entirely: Averaging the two hospitals' weights the ordinary way gives

$$W_{\mathrm{av}}{}^{\mathrm{G}} = \tfrac{1}{2}(1 + (-1)) = 0 \Rightarrow \lambda(T_{\mathrm{av}}{}^{\mathrm{G}}) = \{0,0,0\} \tag{18}$$

a completely degenerate operator. Neither hospital's shared model can propagate information across its own data graph at all, even though each hospital's own weight was individually meaningful. This is destructive interference in its starkest possible form.

GGRS: Propagation restored: Attenuating hospital 2's conflicting update by $\beta = 0.5$, as GGRS's directional soft-weighting would do for a strongly opposed update, gives instead.

$$W^{\mathrm{GGrs}} = \tfrac{1}{2}(1 - 0.5) = 0.25 \Rightarrow \lambda(T^{\mathrm{GGrs}}) = 0.25 \cdot \lambda(\tilde{A}) \tag{19}$$

a coherent, non-degenerate operator that still propagates information for both hospitals, scaled down but never destroyed. This toy example makes visible, in closed form, exactly what the main experiments show empirically at scale: unregulated aggregation can erase a shared model's ability to reason over structure entirely, while GGRS's regulation keeps it alive.