Topology-Aware Differential Privacy in Federated Learning

Murtaza Rangwala, *Member, IEEE,* Richard O. Sinnott, and Rajkumar Buyya, *Fellow, IEEE*

***Abstract*—Federated learning transmits only model updates to protect client data, and differentially private SGD (DP-SGD) bounds content-level leakage through those updates. Neither mechanism accounts for what the communication topology of the federation itself reveals. In cross-silo deployments, a passive adversary with knowledge of the topology and organisational structure has access to information channels that DP-SGD leaves entirely unaddressed. We formalise this threat and derive a principled defense. We introduce TADI (Topology-Aware Distributional Inference), a shadow-trained channel decomposition that isolates per-client leakage into parameter, structural, and organisational components via four channel ablations, and prove an additive per-client mutual-information bound separating a controllable mechanism term from an uncontrollable prior-coupling floor. From this bound we derive FULCRUM, a closed-form balanced min-max optimal noise allocation that strictly dominates uniform DP-SGD whenever the federation's leverage profile is asymmetric, and degenerates exactly to uniform DP-SGD when it is not, making it safe to adopt unconditionally. Evaluated on Fed-ISIC2019, Fed-Heart-Disease, and synthetic CIFAR-10 across six topology families, FULCRUM delivers privacy gains of up to 1.967 nats at no measurable utility cost. The TADI channel decomposition confirms that the parameter channel is bounded by DP-SGD across all settings, the prior-coupling channel is empirically attained under matched-prior conditions, and the bound is conservative in a deployment-favourable direction under realistic cross-silo threat models.**

***Index Terms*—Federated Learning, Differential Privacy, Network Topology**

I. INTRODUCTION

Federated learning (FL) enables collaborative model training across distributed participants without centralising raw data. Clients share only model updates, so sensitive records, whether patient files, clinical measurements, or proprietary assay results, remain local to each participant [1], [2], [3], [4]. The privacy premise is that an adversary observing these updates learns little about the underlying data.

Considerable work has shown this premise requires active enforcement. Gradient inversion attacks can reconstruct individual training samples from unprotected gradient observations [5], [6], membership inference attacks can determine whether a specific record participated in training [7], and property inference attacks can recover distributional attributes of a client's local dataset from the joint update sequence [8]. The standard defense against this class of content-level leakage is differentially private SGD (DP-SGD) [9], which clips and noise-perturbs each client's gradient before transmission, bounding the influence of any individual training record on the observable update. Its federated instantiation, DP-FedAvg [10], applies this mechanism within the FedAvg aggregation protocol, and Renyi differential privacy (RDP) [11] provides the tightest known composition accounting for the cumulative privacy cost across training rounds.

These defenses operate on the content of updates. They do not account for what the communication topology of the federation itself reveals. Real deployments are structurally heterogeneous: hospital consortia route updates through regional aggregators before reaching a central server [12], [13], industrial deployments employ peer-to-peer gossip protocols [14], [15], and various DAG-based aggregation schemes have been proposed for operational scalability [16]. Topology is treated as a performance design parameter [17], and its privacy implications have received little systematic attention. A passive adversary who knows the communication topology and organisational structure of the federation, and who observes per-round parameter updates, has access to two information channels that DP-SGD leaves entirely unaddressed: the structural position of each client in the communication graph, and the organisational membership encoded by the deployment. Together, these channels may expose each client's sensitive-class concentration even when the parameter channel is effectively noise-saturated.

This paper formalises that leakage and derives a principled defense. The information an adversary can extract about a client's data separates additively into two terms: a controllable mechanism term that flows through the parameter updates and shrinks as per-client DP-SGD noise increases, and an uncontrollable prior-coupling term that reflects the client's structural position in the federation and cannot be suppressed by any noise allocation. Standard DP-SGD ignores this asymmetry by applying a uniform noise multiplier across all clients, which is suboptimal whenever the federation topology is asymmetric. We formalise this observation and derive FULCRUM, a closed-form balanced min-max optimal noise allocation that, under a fixed utility budget, strictly dominates uniform DP-SGD whenever structural-leverage scores are non-uniform across clients. When the topology is symmetric and leverage scores are equal, the allocation degenerates to uniform DP-SGD, so it can be adopted unconditionally without penalty.

To ground the decomposition empirically, we introduce TADI (Topology-Aware Distributional Inference), a shadow-trained regressor with four channel ablations that measures the marginal information contributed by the parameter, structural, and organisational channels respectively. TADI is not proposed as a practical attack to be deployed against real federations; rather, it is the empirical instrument through which we verify that the theoretical channel boundaries are meaningful and that the uncontrollable prior-coupling term is realisable when adversary and deployment priors align. The channel decomposition also provides the threat model against

Murtaza Rangwala, Richard O. Sinnott and Rajkumar Buyya are with the Quantum Cloud Computing and Distributed Systems (qCLOUDS) Lab, School of Computing and Information Systems, The University of Melbourne, Australia.

which FULCRUM's guarantees are stated.

The key contributions of this paper are:

1) A formal threat model for topology-conditional distributional inference against DP-protected FL, and an additive per-client mutual-information bound (Theorem V.2) separating a controllable mechanism term from an uncontrollable prior-coupling floor (Sections III–V).
2) The closed-form balanced min-max noise allocation FULCRUM (Theorem V.3), with three principled leverage proxies covering the hierarchical, decentralised, and cross-silo deployment regimes, which strictly dominates uniform DP-SGD whenever structural leverage is asymmetric and degenerates exactly to uniform DP-SGD when it is not (Section V).
3) The TADI channel decomposition with four ablations and a constant-mean baseline comparator, providing the first per-channel empirical characterisation of topology-conditional leakage in DP-FL and confirming the additive structure of Theorem V.2 (Section IV).
4) End-to-end empirical validation on Fed-ISIC2019 [18], Fed-Heart-Disease [18], and synthetic CIFAR-10 [19] with parametric topology-data coupling (Section VI). FULCRUM strictly dominates uniform DP-SGD on the privacy bound at every tested budget and observation window, with no measurable cost to model utility. The parameter channel is bounded by DP-SGD across all settings, the prior-coupling channel is empirically attained when shadow and target priors match, and the bound is conservative in a deployment-favourable direction on real cross-silo deployments.

II. RELATED WORK

We position our work along three axes: privacy attacks against FL, privacy mechanisms for FL, and the largely separate literature on topology in FL. The gap our work fills, namely topology-conditional distributional inference under DP with a defense paired by construction, sits at the intersection of all three and, to our knowledge, has not been addressed.

A. Privacy Attacks Against Federated Learning

Two attack families dominate the literature, both content-oriented. The first, *gradient inversion*, reconstructs training samples by optimising dummy inputs to match observed gradients. Zhu et al. [5] introduced Deep Leakage from Gradients (DLG), which recovers pixel-accurate images from per-batch gradients in the absence of noise. Zhao et al. [20] improved upon this with iDLG, which analytically extracts ground-truth labels before optimisation, and Geiping et al. [6] subsequently demonstrated that magnitude-invariant formulations can recover images even from federated-averaging updates covering multiple local steps. Yin et al. [21] further showed that high-resolution reconstruction remains feasible even with batch sizes larger than those assumed by earlier work. This attack family degrades sharply under DP-SGD because gradient clipping and Gaussian perturbation destroy the high-frequency signal required for pixel-level reconstruction [22]. We treat gradient inversion as a complementary threat relevant primarily when DP noise is weak; it is outside the scope of our empirical evaluation.

The second family, *inference attacks*, targets distributional or membership properties rather than per-sample reconstruction. Shokri et al. [7] established the shadow-model paradigm for membership inference, in which the adversary trains an attack model on shadow datasets where membership is known and applies it to the target. Nasr et al. [23] extended this paradigm to the federated setting, demonstrating that a malicious participant or aggregator can mount membership inference attacks by exploiting gradient updates during training. Melis et al. [8] showed that joint update sequences leak unintended distributional properties of client datasets. Our TADI channel decomposition builds on this shadow-training framework but differs in two respects: it targets a per-client distributional quantity, namely the fraction of each client's data belonging to a sensitive class, rather than membership, and it explicitly decomposes its information sources into parameter, structural-position, and organisational-label channels. To our knowledge, no prior work isolates or quantifies the contribution of these structural channels.

B. Privacy Mechanisms for Federated Learning

The dominant defense against content-level leakage in federated learning is differential privacy applied at the client. Abadi et al. [9] introduced DP-SGD, which clips per-record gradients and adds calibrated Gaussian noise before transmission. McMahan et al. [10] instantiated this mechanism within federated averaging as DP-FedAvg and demonstrated its viability for production-scale next-word prediction. Wei et al. [24] provided the first comprehensive convergence analysis of DP-FL via the noising-before-aggregation framework, which informs our utility cost analysis in Section V. Privacy accounting across training rounds is performed via Renyi differential privacy (RDP) [11], a moment-based composition framework that tracks cumulative privacy cost more tightly than classical bounds. Subsampling amplification [25] extends this accounting to the Poisson-batched setting that DP-SGD requires, where each client samples a random subset of its data per round. The mutual-information (MI) characterisation of DP by Cuff and Yu [26] and its tighter conversions between RDP and other variants [27] provide the machinery we use to derive the per-client MI bound of Theorem V.2.

Beyond noise-based mechanisms, a complementary cryptographic defense exists in the form of secure aggregation. Bonawitz et al. [28] proposed a protocol that masks individual client updates so that only their aggregate is visible to the server. Combined with DP, secure aggregation reduces the adversary's effective observation to neighbourhood-aggregated updates rather than per-client ones, weakening TADI's parameter channel without addressing the structural and organisational channels. FULCRUM's allocation is compatible with secure aggregation as a complementary defense; we discuss the implications in Section VII and leave a full empirical evaluation under secure aggregation to future work.

A separate line of work permits per-individual privacy budgets driven by user preferences, rather than a uniform budget

applied across all clients. Jorgensen et al. [29] introduced personalised differential privacy for sample-release queries, and subsequent work has extended this to ML training. Our setting is distinct: the asymmetry that motivates per-client noise allocation is structural, arising from topology and organisational position, rather than preferential, and the allocation is derived from a worst-case MI bound rather than from declared user preferences. The two lines are complementary, and an integrated allocation combining structural leverage with per-user preferences is a natural direction that falls outside the scope of this work.

What none of these mechanisms address is which clients should receive how much noise when the topology breaks symmetry. Existing implementations apply a uniform noise multiplier across all clients, implicitly assuming a symmetric topology. Our defense generalises this: when the topology is symmetric, the allocation reduces to uniform DP-SGD; when it is not, the allocation balances the per-client privacy bound across the asymmetric leverage profile.

C. Topology in Federated Learning

As discussed in Section I, topology-aware FL has been studied primarily as a performance and scalability question. The earliest cross-device benchmark suite, LEAF [30], demonstrated that natural client-partitioned datasets exhibit substantial heterogeneity but did not consider topology beyond the implicit star aggregator. Hierarchical FL [12], [13] introduced edge aggregators between clients and the cloud to reduce communication cost. Decentralised gossip-based schemes [14], [15] eliminate the central aggregator entirely, and Beilharz et al. [16] proposed a DAG-based extension for further operational flexibility. Briggs et al. [31] cluster clients by local-update similarity to address non-IID data, producing a hierarchical training structure that incidentally encodes data similarity. The field guide survey by Wang et al. [17] documents topology choice as one of the canonical decisions in FL system design but treats it as orthogonal to privacy.

The privacy literature has largely assumed away topology. Threat models almost universally assume a star configuration with a single semi-honest aggregator, and the attacker's view, when explicitly modelled, consists of per-client gradients without structural metadata. This assumption is reasonable for cross-device deployments where every client is interchangeable, but it breaks down in cross-silo healthcare [18], [1], [2], multi-organisation industry consortia [3], and any deployment that uses hierarchical or decentralised aggregation for operational reasons.

A notable exception is the line of work on topology as a privacy amplification mechanism. Cyffers and Bellet [32] showed that in fully decentralised algorithms, the restricted local view of each participant provides stronger privacy guarantees than local differential privacy alone. This perspective treats topology as a privacy asset for honest participants. Our work takes the opposite view: we assume the adversary has global knowledge of the topology and study how structural position and organisational membership become information channels for inferring sensitive data properties. Feng et al. [33] occupy a third position, demonstrating that the topology itself can be inferred from model behaviour alone. Our threat model differs from both: we assume topology is known and treat it as a leakage surface rather than an amplification mechanism or a reconstruction target.

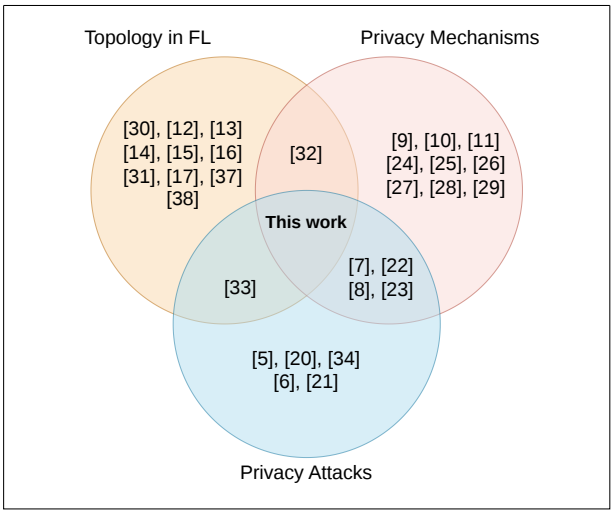

Fig. 1. Positioning of this work relative to the three bodies of literature it spans. Prior work addresses at most two of the three axes; this paper is the first to operate at their intersection.

No prior work treats topology as a known information channel that an adversary can exploit to infer sensitive data properties, nor derives a defense that accounts for the resulting structural asymmetry. We are the first to define a per-client distributional channel decomposition with explicit ablations isolating parameter, structural, and organisational information; prove an additive per-client MI bound separating a controllable mechanism term from an uncontrollable prior-coupling floor; and derive a closed-form noise allocation that strictly improves on uniform DP-SGD whenever the topology breaks leverage symmetry. Figure 1 summarises this positioning relative to the three bodies of literature surveyed above.

III. System and Threat Model

We consider a passive, honest-but-curious adversary against a DP-protected federated learning system, reflecting the realistic scenario of a curious infrastructure observer or aggregator who follows the protocol but exploits what they can passively observe. Active protocol manipulation, malicious aggregators, and Sybil attacks are outside our scope and discussed in Section VII. The channel decomposition and defense are grounded in the same formal assumptions, ensuring that FULCRUM is evaluated against precisely the adversary it is designed to resist.

A. System Model

A federation consists of n clients $\mathcal{P} = \{P_1, \ldots, P_n\}$ arranged in a communication topology $\mathcal{G} = (\mathcal{P}, E)$, where

E denotes the set of edges determining which pairs of clients communicate during aggregation. Each client P_i holds a local dataset $\mathcal{D}_i$ over a class set $\mathcal{C}$ of size $K = |\mathcal{C}|$, inducing a local class distribution $\Delta_i \in \Delta(\mathcal{C})$. Clients are partitioned into organisational groups by a labelling $\omega : \mathcal{P} \to [k_{\text{org}}]$, where k_{org} is the number of distinct groups. Training proceeds for $T_{\max}$ rounds. At each round t, client P_i applies local DP-SGD with per-client noise scale σ_i to produce an updated parameter vector $\theta_i^{(t)}$, which is then transmitted to its neighbours in $\mathcal{G}$. Topologies are assumed to be static throughout training. Client datasets are assumed to be disjoint: no two clients share training records. This is the standard cross-silo assumption [17], satisfied naturally in settings such as hospital consortia, where each site enrols distinct patients. Aggregation follows FedAvg [10] for centralised hierarchical settings and decentralised gossip [15] for peer-to-peer topologies; both reduce to weighted averages of client parameter vectors.

B. Adversarial Model

The adversary $\mathcal{A}$ is specified by four components: knowledge $\mathcal{K}$, observations $\mathcal{O}$, an inference goal $\mathcal{I}$, and a set of restrictions $\mathcal{R}$.

The adversary's knowledge $\mathcal{K}$ comprises the communication topology $\mathcal{G}$, the organisational labelling ω, and all public protocol parameters, including the DP budget ε, the per-client noise schedule $\{\sigma_j\}$, the model architecture, and the training schedule. The adversary does not know any client's local dataset $\mathcal{D}_i$ or its induced class distribution Δ_i. We treat $\mathcal{K}$ as a configuration constant, conditioning all probabilistic claims on it throughout.

The adversary's observations $\mathcal{O}$ consist of the full tensor of transmitted parameter updates, $\Theta = \{\theta_i^{(t)} : i \in [n], t \in [T_{\max}]\}$, together with communication metadata. This corresponds to a passive infrastructure observer, such as a curious cloud aggregator or a network-level observer with access to unencrypted traffic.

The inference goal $\mathcal{I}$ is to recover, for each client i, the sensitive-class concentration

$$p_i \;=\; \Delta_i(\mathcal{C}_s) \;\in\; [0,1],$$

where $\mathcal{C}_s \subset \mathcal{C}$ is a designated sensitive class set. This captures either a single sensitive proportion in binary classification tasks, or a designated rare-class proportion in multiclass settings. In cross-silo deployments, p_i encodes population-level attributes that organisations are legally and ethically obligated to protect [1], [2]: the melanoma rate per hospital site, the cardiovascular disease prevalence per clinical centre, or the failure rate of a proprietary assay. An adversary who recovers these concentrations across sites can exploit them for adversarial site selection, competitive intelligence, or regulatory disclosure inference [34], harms that are concretely realised in the deployment contexts federated learning was designed to serve. Inference performance is measured by calibration loss $L_{\text{cal}} = \frac{1}{n}\sum_i (\hat{p}_i - p_i)^2$ and channel lift relative to the constant-mean baseline $\bar{p} = \frac{1}{n}\sum_j p_j$, which measures whether a given channel contributes information beyond the federation aggregate.

The restrictions $\mathcal{R}$ define an honest-but-curious adversary. The adversary follows the protocol faithfully, cannot perturb $\theta_i^{(t)}$ prior to observation, cannot inject training data, cannot collude with clients, and cannot adaptively select targets.

C. Independence Assumptions

Two assumptions underpin the formal analysis.

(IA1) DP-SGD noise independence. The per-round Gaussian noise $\xi_i^{(t)} \sim \mathcal{N}(0, \sigma_i^2 C^2 I)$ is independent across all (i,t) pairs. This is standard in the DP-SGD literature [9].

(IA2) Disjoint client datasets. For $i \neq j$, $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset$, as established in Section III-A. Cross-device settings with overlapping client datasets would require strengthened conditions; see Section VII.

Aggregation follows FedAvg or its hierarchical and gossip variants and is not treated as a privacy mechanism in our threat model, since the adversary observes per-client updates prior to aggregation. Under secure aggregation [28] this changes; FULCRUM is compatible with secure aggregation as a complementary defense, and we discuss this in Section VII.

D. Reference Adversaries

To isolate the contribution of each information channel, we define five reference adversaries differing only in their knowledge $\mathcal{K}$, with observations $\mathcal{O} = \Theta$ in all cases.

- $\mathcal{A}_0$ (constant-mean baseline): outputs the federation mean $\bar{p}$ for every client. Serves as the reference for computing channel lift.
- $\mathcal{A}_1$ (parameter-only): $\mathcal{K} = \{\sigma_j\}$. Recovers $\hat{p}_i$ from parameter sequences alone, with no access to structural or organisational information. This is the topology-agnostic adversary of Melis et al. [8].
- $\mathcal{A}_2^{\text{topo}}$ (structural): $\mathcal{K} = \mathcal{G}$. Augments parameter sequences with structural features including degree, hierarchical depth, and distance to root, isolating the marginal information value of the communication topology.
- $\mathcal{A}_2^{\text{org}}$ (organisational): $\mathcal{K} = \omega$. Augments parameter sequences with organisational labels, isolating the marginal information value of organisational membership.
- $\mathcal{A}_2^{\text{full}}$ (full knowledge): $\mathcal{K} = (\mathcal{G}, \omega, \{\sigma_j\})$. The primary adversary of interest.

The topology contribution is measured as $L_{\text{cal}}(\mathcal{A}_1) - L_{\text{cal}}(\mathcal{A}_2^{\text{full}})$, and the marginal contributions of structure and organisational label are obtained by comparing $\mathcal{A}_2^{\text{full}}$ against the corresponding partial adversary. The IID-null condition arises when the data partitioning makes p_i constant across clients, in which case every adversary's calibration loss equals $\text{Var}(p)$ and channel lift is identically zero. We use this condition as a consistency check that holds by construction throughout the empirical evaluation.

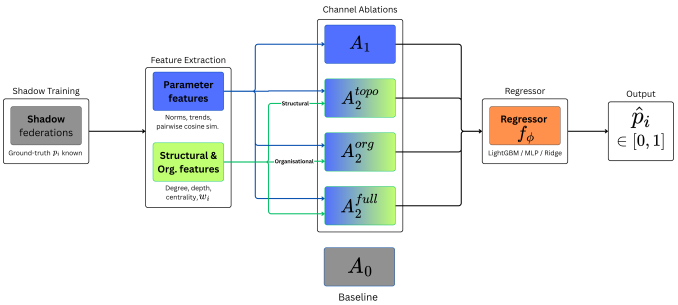

Fig. 2. The TADI channel decomposition pipeline. Shadow federations are simulated to harvest labelled training tuples, parameter and structural features are extracted, a channel subset corresponding to one of the four ablations is selected, and a regressor f_ϕ is fitted to estimate the sensitive-class concentration $\hat{p}_i$ for each client. $\mathcal{A}_0$ is the constant-mean baseline and does not correspond to a channel ablation.

IV. The TADI Channel Decomposition

TADI is a learned regressor with four channel ablations, each corresponding to a different subset of the adversary's available information. The architecture maps observable inputs to per-client sensitive-class concentration estimates, and the ablations differ only in which channels of $\mathcal{K}$ and $\mathcal{O}$ are exposed to the regressor. This design serves two purposes: it provides the empirical characterisation of topology-conditional leakage, and it ensures the channel boundaries align with the decomposition required by the theoretical bound in Section V.

Let $f_\phi : (\Theta_i, x_i) \to \hat{p}_i \in [0,1]$ denote the regressor with learnable parameters ϕ. The regressor takes two classes of input, as illustrated in Figure 2. The first is the parameter sequence $\Theta_i = (\theta_i^{(1)}, \ldots, \theta_i^{(T_{\max})})$, from which we extract a fixed-dimensional feature vector combining per-round and temporal-aggregate signals. Per-round signals comprise parameter norms $\{\|\theta_i^{(t)}\|_2\}$, layer-wise norms, and pairwise cosine similarities to neighbouring clients' updates. Temporal aggregates comprise the mean and standard deviation of parameter norms across rounds, a linear trend estimate, and the last-round parameter value. The second is a structural and organisational feature vector $x_i \in \mathbb{R}^k$ derived from $(\mathcal{G}, \omega)$, comprising client i's degree in $\mathcal{G}$, its distance to the topology root, and its betweenness centrality, with the organisational label ω_i encoded as a one-hot vector and concatenated. Distance to root is defined for rooted topologies such as hierarchical FL; for decentralised peer-to-peer topologies without a natural root, this feature is set to zero. Betweenness centrality is precomputed from the known topology $\mathcal{G}$ prior to training.

Given these two input classes, the regressor is trained offline via the shadow-model framework [7]. Shadow federations are simulated on a proxy dataset where the ground-truth p_i is known, (Θ_i, x_i, p_i) tuples are harvested across many simulated clients, and f_ϕ is fitted to minimise mean squared calibration loss on the harvested set. At test time, f_ϕ is applied to the target federation. The regressor is therefore a deterministic function of $(\mathcal{K}, \mathcal{O})$, consistent with the threat model of Section III-B. The regressor is instantiated as gradient-boosted trees (LightGBM, default), a multilayer perceptron, or ridge regression, all clipped to $[0,1]$. Algorithm 1 gives the formal specification.

Algorithm 1 TADI, full-channel variant ($\mathcal{A}_2^{\text{full}}$)

Require: Shadow corpus $\mathcal{S} = \{(\Theta_i, x_i, p_i)\}_{i=1}^N$, target observations $(\Theta^{\text{tgt}}, \mathcal{G}, \omega, \{\sigma_j\})$, regressor backend $\mathcal{B}$
Ensure: $\{\hat{p}_i\}_{i=1}^n$

$\phi_\Theta^{(i)} \leftarrow \textsc{ParamFeatures}(\Theta_i)\ \forall (\Theta_i, x_i, p_i) \in \mathcal{S}$
$\phi_x^{(i)} \leftarrow [\deg_{\mathcal{G}}(i),\ \text{depth}(i),\ \text{cent}(i),\ \text{onehot}(\omega_i)]\ \forall i$
$f \leftarrow \mathcal{B}\big(\{[\phi_\Theta^{(i)}; \phi_x^{(i)}]\}_i, \{p_i\}_i\big)$
for $i = 1$ **to** n **do**
  $\phi_\Theta^{(\text{tgt},i)} \leftarrow \textsc{ParamFeatures}(\Theta_i^{\text{tgt}})$
  $\hat{p}_i \leftarrow \text{clip}_{[0,1]}\big(f([\phi_\Theta^{(\text{tgt},i)}; \phi_x^{(\text{tgt},i)}])\big)$
end for
return $\{\hat{p}_i\}_{i=1}^n$

To isolate the contribution of each information source, the four channel ablations correspond to the reference adversaries of Section III-D and differ only in which features are exposed to the regressor. $\mathcal{A}_1$ receives parameter features only. $\mathcal{A}_2^{\text{topo}}$ receives parameter features augmented with structural features, comprising degree, depth, and centrality. $\mathcal{A}_2^{\text{org}}$ receives parameter features augmented with the one-hot organisational label ω_i. $\mathcal{A}_2^{\text{full}}$ receives all three concatenated. A separate regressor is trained per channel on the same shadow corpus, ensuring that each $\hat{p}_i$ represents the best the regressor can achieve given exactly that channel's input.

Performance across the four ablations is evaluated using the metrics summarised in Table I. Here top-$k(\hat{p})$ and top-$k(p)$ denote the sets of k clients with the highest predicted and true sensitive-class concentration respectively, and τ is a concentration threshold used to binarise the ranking task for AUROC.

The channel decomposition is designed to answer three empirical questions. First, does DP-SGD effectively bound the parameter channel under the per-client noise allocation, independently of any prior coupling? Second, which non-parameter channel, structural position or organisational label, carries the prior-coupling signal in a given deployment? Third, is the prior-coupling signal realisable under the public-proxy threat model, or does the gap between the adversary's shadow prior and the target's deployment prior place the supremum ℓ_i° out of empirical reach? These questions are addressed in Section VI.

V. Defense: Topology-Aware DP Allocation

The leakage an adversary can extract about a client's data separates additively into a mechanism term that flows through parameter updates and can be suppressed by noise, and a prior-coupling term that flows through structural position and cannot. The defense targets the controllable component by allocating the per-client noise scale σ_i asymmetrically as a function of structural-leverage scores, rather than applying a uniform multiplier across all clients as standard practice assumes. We prove a per-client mutual-information bound that holds against the adversary of Section III-B and derive Fulcrum, a closed-form balanced min-max optimal allocation that strictly dominates uniform DP-SGD whenever leverage

TABLE I
CHANNEL DECOMPOSITION EVALUATION METRICS COMPUTED PER TARGET RUN AND CHANNEL.

Metric	Definition	Notes
Calibration loss	$L_{\text{cal}}(\hat{p}, p) = \frac{1}{n}\sum_i (\hat{p}_i - p_i)^2$	Primary metric; retains statistical power at small n
Constant-mean baseline	$L_{\text{cal}}(\bar{p}, p) = \text{Var}(p)$	Trivial predictor; reference for computing channel lift
Channel lift	$L_{\text{cal}}(\bar{p}, p) - L_{\text{cal}}(\hat{p}, p)$	Positive iff channel beats baseline; zero under IID-null
Top-k recovery	$\lvert\text{top-}k(\hat{p}) \cap \text{top-}k(p)\rvert / k$	Bounded, interpretable ranking metric
AUROC	For $y_i = \mathbb{1}[p_i > \tau]$	Ranking quality indicator; meaningful at $n \geq 20$

scores are non-uniform across clients, and degenerates exactly to uniform DP-SGD when they are not.

A. Notation and Setup

The world prior $\mathbb{P}_{\text{world}}$ generates client datasets $D = (D_1, \ldots, D_n)$ with class distributions $\Delta_i = \delta(D_i)$ and sensitive-class concentrations $p_i = \Delta_i(\mathcal{C}_s)$. DP-SGD [9] produces, per round and per client,

$$\theta_i^{(t)} = \theta^{(t-1)} - \alpha \cdot \tilde{g}_i^{(t)},$$
$$\tilde{g}_i^{(t)} = \frac{1}{|B|} \sum_{x \in B_i^{(t)}} \text{clip}\big(\nabla \ell(\theta^{(t-1)}, x), C\big) + \xi_i^{(t)}, \qquad (1)$$

where α is the learning rate, $B_i^{(t)}$ is the mini-batch sampled by client i at round t, ℓ is the loss function, and $\xi_i^{(t)} \sim \mathcal{N}(0, \sigma_i^2 C^2 I)$ is the per-round Gaussian noise introduced by DP-SGD, where $\sigma_i > 0$ is a dimensionless noise multiplier following Abadi et al. [9]: the noise is added to the per-sample average gradient, so the effective noise scale on the averaged gradient is $\sigma_i C$, and the sensitivity of the averaged gradient is $C/|B|$. The adversary observes $\Theta = \{\theta_i^{(t)}\}_{i,t}$ and outputs estimate $\hat{p}_i$ via a deterministic function f as defined in Section IV. We let $T_{\max}$ denote the observation window length, that is, the number of rounds visible to the adversary, which may be capped if a windowing or re-keying defense is in place.

The analysis is constrained to a family of priors $\mathcal{F}_{\mathcal{G},\omega}$ over $(\Delta_1, \ldots, \Delta_n)$ satisfying three conditions: (a) the prior factorises over connected components of $\mathcal{G}$; (b) within each component, the prior satisfies a Markov property with respect to $\mathcal{G}$; and (c) the prior has bounded second moments. These conditions are standard in graphical-model analyses. Condition (c) ensures that the supremum quantities introduced below are finite, and all three conditions are satisfied by the deployment priors used in Section VI.

B. Structural Leverage

The key quantity driving the noise allocation is the degree to which a client's structural position in the federation makes its sensitive-class concentration p_i statistically predictable from the rest of the federation's data. A client at a structurally exposed position, such as the hub of a star topology or the sole member of a small organisational group, carries more lateral information about p_i than a client in a symmetric or isolated position. We formalise this intuition through the concept of structural leverage.

Definition V.1 (Structural leverage)**.** *For client i in deployment $(\mathcal{G}, \omega)$, the structural leverage is*

$$\ell_i^\circ := \sup_{\mathbb{P} \in \mathcal{F}_{\mathcal{G},\omega}} I_{\mathbb{P}}\big(p_i;\, D_{-i}\big),$$

where $D_{-i} = (D_1, \ldots, D_{i-1}, D_{i+1}, \ldots, D_n)$ denotes the datasets of all clients other than i.

The supremum captures the worst-case prior coupling between p_i and the rest of the federation's data, taken over all priors consistent with the deployment structure. It is a deterministic function of $(\mathcal{G}, \omega, i)$, finite under condition (c) of Section V-A, and computable in closed form for the practical proxies introduced in Section V-E. Clients with higher structural leverage are more exposed to distributional inference and therefore warrant stronger noise protection under the allocation derived in Section V-D.

C. Per-Client Conditional MI Bound

With structural leverage defined, we can now bound the mutual information an adversary can extract about any individual client. The bound decomposes additively into the two terms identified informally in Section I: a controllable mechanism term that shrinks with increasing per-client noise, and an uncontrollable prior-coupling floor determined by the client's structural position.

Theorem V.2 (Per-client conditional MI bound)**.** *Under (IA1) and (IA2), for any deterministic adversary strategy f, any prior $\mathbb{P} \in \mathcal{F}_{\mathcal{G},\omega}$, and any client i:*

$$I_{\mathbb{P}}(p_i; \hat{p}_i \,|\, \mathcal{G}, \omega, \{\sigma_j\}) \;\leq\; \underbrace{\frac{T_{\max}}{2\sigma_i^2 |B|^2}}_{\textit{controllable}} \;+\; \underbrace{\ell_i^\circ}_{\textit{uncontrollable}}. \qquad (2)$$

Proof sketch. By the data processing inequality, $\hat{p}_i = f(\Theta)$ implies $I(p_i; \hat{p}_i) \leq I(p_i; \Theta)$. Applying the chain rule and dropping the non-negative term $I(p_i; D_{-i} \mid \Theta) \geq 0$ gives $I(p_i; \Theta) \leq I(p_i; \Theta \mid D_{-i}) + I(p_i; D_{-i})$. Under (IA2), the first term reduces to $I(D_i; \Theta_i \mid D_{-i})$ and is bounded by $T_{\max}/(2\sigma_i^2|B|^2)$ via Mironov's RDP composition [11] applied to the Gaussian mechanism [9], converted to mutual information via the max-KL bound of Cuff and Yu [26]. The second term is bounded by ℓ_i° by Definition V.1. The full proof is given in the supplementary material (Appendix A). □

The additive structure is the central feature of the bound. The first term is controllable: it shrinks as the per-client noise σ_i increases and is the target of the allocation derived in Section V-D. The second term is uncontrollable: it is a property

of the deployment prior and cannot be reduced by any amount of DP-SGD noise applied to client i. The allocation therefore manages the controllable term given that the prior-coupling floor is fixed.

Theorem V.2 is an upper bound rather than a tight equality, and two sources of slack are worth noting. First, the chain-rule step drops the non-negative term $I(p_i; D_{-i} \mid \Theta) \geq 0$, introducing conservatism. Second, the Cuff-Yu max-KL conversion [26] is loose at extreme noise levels. In the implementation of Section VI, Poisson subsampling at rate $q = |B|/|\mathcal{D}_i|$ is applied via Opacus [35], under which amplified RDP [25] would tighten the per-round contribution from $1/(2\sigma_i^2|B|^2)$ to $1/(2\sigma_i^2|\mathcal{D}_i|^2)$, a substantial improvement when $|\mathcal{D}_i| \gg |B|$. Tighter bounds via Asoodeh et al. [27] or amplified RDP are natural extensions of the proof structure and are discussed in Section VII.

D. Balanced Min-Max Optimal Allocation

Given the per-client MI bound of Theorem V.2, the natural objective is to minimise the worst-case bound across all clients subject to a fixed total noise budget. Let $a := T_{\max}/(2|B|^2)$. The optimisation problem is

$$\min_{\{\sigma_i^2 > 0\}} \max_i \left[a/\sigma_i^2 + \ell_i^\circ \right] \quad \text{subject to} \quad \sum_{i=1}^n \sigma_i^2 \leq U, \tag{3}$$

where U is the total noise variance budget. The total noise variance $\sum_i \sigma_i^2$ governs the utility cost of DP-SGD under standard FL convergence assumptions [24], justifying U as a utility budget proxy and motivating the constraint in (3). Intuitively, increasing U admits larger per-client noise scales and therefore higher utility cost. The following theorem shows that (3) admits a closed-form solution.

Theorem V.3 (Balanced min-max allocation)**.** *For any utility budget $U > 0$, the optimal allocation solving* (3) *is*

$$\sigma_i^{*\,2} = \frac{a}{K^\star - \ell_i^\circ}, \tag{4}$$

where $K^\star$ is the unique solution to

$$\sum_{i=1}^n \frac{a}{K^\star - \ell_i^\circ} = U, \quad K^\star > \max_i \ell_i^\circ. \tag{5}$$

The achieved worst-case per-client MI bound is $K^\star$, equilibrated uniformly across clients.

Proof sketch. Reformulate (3) via a slack variable K: minimise K subject to $a/\sigma_i^2 + \ell_i^\circ \leq K$ for all i and $\sum_i \sigma_i^2 \leq U$. The objective is linear in K, the constraints are convex since a/x is convex on $x > 0$, and Slater's condition holds at $\sigma_i^2 = U/(2n)$. Karush-Kuhn-Tucker (KKT) conditions therefore give $\sigma_i^{*\,2} = a/(K^\star - \ell_i^\circ)$ at the active-budget optimum; uniqueness of $K^\star$ follows from strict monotonicity of the budget equation in K. The full proof is given in the supplementary material (Appendix B). □

An immediate consequence of Theorem V.3 is that the topology-aware allocation strictly dominates uniform DP-SGD whenever the federation is asymmetric.

Corollary V.4 (Strict improvement over uniform allocation)**.** *Under uniform allocation $\sigma_i^2 = U/n$, the worst-case bound is $K_{\text{uniform}} = an/U + \max_i \ell_i^\circ$. The balanced allocation of Theorem V.3 satisfies $K^\star \leq K_{\text{uniform}}$, with equality if and only if all ℓ_i° are equal.*

Corollary V.4 confirms that the topology-aware allocation improves on uniform DP-SGD exactly when it should. Symmetric topologies such as rings, complete graphs, and balanced hierarchies produce uniform leverage scores, and the allocation reduces to standard uniform DP-SGD. This is the appropriate behaviour: there is no asymmetry to exploit and no penalty for using FULCRUM unconditionally. Asymmetric topologies such as lines, hierarchies with uneven groups, and stars, as well as deployments with heterogeneous client dataset sizes, produce non-uniform leverage and admit a strict improvement.

E. Practical Leverage Proxies

The abstract leverage ℓ_i° is well-defined but not always analytically computable in closed form. We instantiate three principled proxies, each capturing a different source of structural asymmetry. Theorems V.2 and V.3 hold for any non-negative weight vector substituted in place of $\{\ell_i^\circ\}$: the allocation strictly dominates uniform DP-SGD whenever the weights are non-uniform, regardless of whether they equal the true structural leverage. The proxies are therefore validated on two criteria: ordering consistency, meaning the proxy correctly ranks clients by structural exposure, which is sufficient for Corollary V.4 to apply; and empirical conservatism, meaning the allocation consistently reduces K_{uniform} in practice. Tight analytical bounding of ℓ_i° from above by each proxy, which would give a guaranteed per-client MI bound, holds for the group-size proxy asymptotically and remains open for the degree and dataset-size proxies.

1. Group-size proxy. Under a stochastic block model prior on $(\Delta_1, \ldots, \Delta_n)$ with $\Delta_i \sim_{\text{iid}} \mathbb{P}_{\Phi_{\omega_i}}$ given finite-entropy block parameter Φ, it can be shown that $\ell_i^\circ \leq H(\Phi_{\omega_i})$, asymptotically independent of block size. For finite blocks, ℓ_i° is approximated by the group size $|G_i|$ scaled by a prior-dependent constant, motivating the proxy $\ell_i^{\text{org}} \propto |G_i|$. This proxy is ordering-consistent under the SBM prior and satisfies the asymptotic bounding condition. It is appropriate whenever the dominant source of asymmetry is uneven organisational groups, as in hierarchical FL with heterogeneous clusters [31].

2. Degree proxy. Under bounded-degree Markov priors with a regularity assumption on neighbourhood-conditional entropy, leverage admits an upper bound proportional to graph degree. We adopt the proxy $\ell_i^{\deg} \propto \deg_{\mathcal{G}}(i)$ as a heuristic for deployments where the dominant asymmetry arises from non-uniform graph degree, such as decentralised peer-to-peer protocols [14], [15], [16]. Ordering consistency holds when degree is the primary driver of neighbourhood entropy; a tight analytical bound requires characterising the conditional entropy at each neighbourhood as a function of global graph structure, which resists a closed-form treatment and remains open. Empirical validation across six topology families, including graphs where degree heterogeneity is the defining structural property, provides the practical basis for its use.

3. Dataset-size proxy. In FedAvg with size-weighted aggregation $\theta^{(t+1)} = \sum_i \frac{|\mathcal{D}_i|}{\sum_j |\mathcal{D}_j|}\theta_i^{(t+1)}$, each client's update is weighted proportionally to its dataset size, making larger clients structurally more influential. Under a Dirichlet-multinomial prior on client class distributions, the lateral mutual information $I(p_i; D_{-i})$ scales with the client's aggregation weight, giving $\ell_i^\circ \propto |\mathcal{D}_i|$ under this generative model. We adopt the mean-normalised proxy $\ell_i^{\mathrm{ds}} \propto |\mathcal{D}_i|/|\bar{\mathcal{D}}|$, which is ordering-consistent under the Dirichlet-FedAvg model by construction. A tight upper bound on ℓ_i° beyond the ordering result remains open, as in the degree proxy case. The Pareto dominance results of Section VI-D provide empirical validation across all tested budgets and observation windows.

The three proxies are not mutually exclusive. A deployment can blend them by passing per-client weight vectors that combine group size, degree, and dataset size, normalised to unit mean. The choice of proxy is a deployment design parameter; each experimental setting in Section VI uses the proxy that matches its dominant source of asymmetry.

VI. Performance Evaluation

We evaluate both contributions on three federated benchmarks spanning the canonical cross-silo deployment regimes. The TADI channel decomposition is evaluated on its ability to characterise sensitive-class concentration leakage across the four ablations, and FULCRUM is evaluated on its privacy-utility trade-off relative to uniform DP-SGD. The benchmarks are chosen to cover different sources of structural asymmetry: heterogeneous site sizes in a real healthcare imaging federation, heterogeneous site sizes in a real cross-silo tabular task, and parametric topology-data coupling in a controlled synthetic setting. Together they provide both deployment realism and the statistical power needed to validate our theoretical claims.

A. Settings and Datasets

Setting A: Fed-ISIC2019. The Fed-ISIC2019 component of FLamby [18] comprises dermoscopy images from six hospital sites with eight skin lesion classes, where melanoma is the designated sensitive class $\mathcal{C}_s$. Site sizes are highly heterogeneous across the six centres, with the largest site contributing nearly thirty times more training samples than the smallest. We arrange the six sites in a two-level hierarchy with three regional aggregators following the architectural pattern of Liu et al. [12], and train a small convolutional network with GroupNorm replacing BatchNorm for DP-SGD compatibility. This setting combines dataset-size asymmetry with hierarchical topology structure, so both the dataset-size proxy and the organisational position of each site within the hierarchy contribute to the leverage profile.

Setting B: Fed-Heart-Disease. The Fed-Heart-Disease component of FLamby [18] provides records across four clinical centres with 13 clinical features and a binary heart-disease indicator. With four clients arranged on a ring topology, all nodes share the same degree and the topology introduces no structural leverage asymmetry. Dataset-size heterogeneity is therefore the sole source of asymmetry, with the largest centre contributing more than six times the training samples of the smallest, making this a clean test of the dataset-size proxy in isolation from any topological confound.

Setting C: Synthetic CIFAR-10 with parametric coupling. To provide controlled statistical power for validating the theoretical claims of Section V, we partition CIFAR-10 [19] across $n = 50$ synthetic clients using the η-coupled construction

$$\Delta_i = (1-\eta)\,\tilde{\Delta}_i + \eta\,\Delta^*_{\phi(i)},$$

where $\tilde{\Delta}_i \sim \mathrm{Dirichlet}(\alpha)$ provides the standard non-IID heterogeneity component [36], $\Delta^*_{\phi(i)}$ is concentrated on class $\phi(i) \bmod K$, and $\eta \in [0,1]$ interpolates between the IID regime ($\eta = 0$) and fully position-determined class distributions ($\eta = 1$). Topology is varied across ring, line, hierarchical, star, Erdős-Rényi (ER) [37], and Barabási-Albert (BA) [38] configurations to expose the full leverage-asymmetry spectrum, from uniform-degree graphs to scale-free networks with highly non-uniform degree distributions. Setting C is the only configuration in which AUROC is a meaningful metric ($n \geq 20$) and provides clean control over η as the parameter driving lateral leakage.

B. Methodology

DP-SGD is implemented via Opacus [35] with Poisson subsampling, the RDP accountant [11], gradient clipping at $C = 1.0$, and batch size $|B| = 64$. The federated training infrastructure builds on the Murmura framework [39]. Each configuration is run with three independent seeds. Per-client noise scales are set either by the topology-aware allocation $\sigma_i^{*2} = a/(K^\star - \ell_i^\circ)$ from Theorem V.3, or uniformly as $\sigma_i^2 = U/n$ for the baseline. The privacy bounds $K^\star$ and K_{uniform} are computed analytically by one-dimensional bisection of the budget and carry no seed variance; test accuracy is reported as the mean over three seeds with 95% confidence intervals.

The defense is evaluated across a factorial sweep of utility budgets U and observation windows $T_{\max}$ under both allocations, with the per-setting sweep grids stated in the relevant subsections below. Utility consistency is assessed by matched pairs at fixed U and seed, giving 18 paired observations per $T_{\max}$. We use Two One-Sided Tests (TOST) at a ± 0.5 percentage-point margin as the primary equivalence test [40], since a TOST p-value below 0.05 actively rejects practically meaningful difference rather than merely failing to reject equality. The conventional paired t-test p-value is reported alongside for completeness.

The TADI channel decomposition is evaluated separately from the defense sweep. One regressor per channel is trained on a setting-specific shadow corpus and applied to all target runs from the Pareto sweep. The shadow corpus construction differs across settings along the prior-matching axis of the threat model. For Setting C, the shadow corpus uses the same η-coupled partitioning as the target, so shadow and target priors match by construction and the prior-coupling supremum ℓ_i° is realisable in principle. For Setting B, the shadow corpus is built from synthetic Dirichlet re-partitioning of the public FLamby data, the only construction available to an adversary without access to native per-site labels, so the

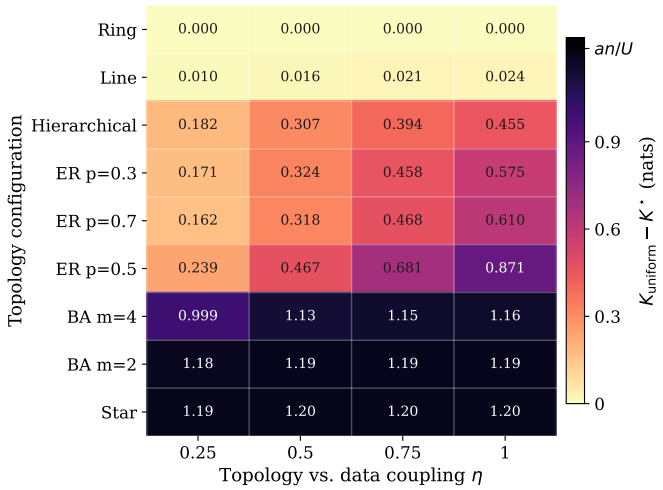

Fig. 3. Privacy-bound gap $K_{\text{uniform}} - K^\star$ (nats) across nine topology configurations and four coupling strengths $\eta \in \{0.25, 0.5, 0.75, 1\}$ on Setting C ($n = 50$, $U = 0.5$, $T_{\max} = 100$). The $\eta = 0$ column is omitted as all values are analytically zero. Rows are sorted by maximum gap. The ring row is identically zero at every η. Scale-free topologies (BA, star) saturate within 2% of the analytic asymptote $an/U \approx 1.22$ nats, marked on the colourbar. Among ER configurations, the gap peaks non-monotonically at $p = 0.5$, where the binomial degree variance is maximised.

shadow and target priors are deliberately mismatched. Setting A is omitted from the channel ablation because its native FLamby partitioning produces the same qualitative pattern as Setting B under the public-proxy threat model; the Pareto sweep and utility consistency results of Section VI-D provide full defense coverage for that setting. Statistical significance across channels is assessed via paired bootstrap on calibration loss with multiple-comparison correction.

C. Theorem Validation: The η-Sweep

The clearest single test of Theorem V.3 and Corollary V.4 is the Setting C η-sweep, which varies the topology-data coupling $\eta \in \{0, 0.25, 0.5, 0.75, 1\}$ across nine topology configurations at fixed $U = 0.5$ and $T_{\max} = 100$. Figure 3 shows the privacy-bound gap $K_{\text{uniform}} - K^\star$ (nats) across the 9×4 grid at $\eta \in \{0.25, 0.5, 0.75, 1\}$.

The heatmap simultaneously confirms five predictions of the theory. At $\eta = 0$, both allocations are analytically equivalent on every topology since no topology-data correlation exists; this column is omitted from Figure 3 as all values are identically zero. The ring row is likewise identically zero at every η: a ring assigns equal degree to all clients, producing a uniform leverage profile under which the allocation degenerates to uniform DP-SGD regardless of coupling strength. Both observations confirm the degenerate case of Corollary V.4.

For every asymmetric topology, the gap grows monotonically in η. The growth is near-linear for moderate-asymmetry configurations, specifically line, hierarchical, and ER, and saturates rapidly for extreme-asymmetry configurations. BA topologies ($m = 2$, $m = 4$) and the star reach approximately 98% of the analytic asymptote $an/U = 1.2207$ nats by $\eta = 0.25$, with the star reaching 1.196 nats and BA $m = 2$ reaching 1.193 nats at $\eta = 1$, both within 2% of the asymptote. Crucially, the star is not uniquely special: any topology with extreme degree concentration saturates the bound. The BA configurations establish this generalisation, extending the worst-case claim from star topologies to scale-free networks more broadly. Among the three ER configurations, the gap peaks non-monotonically at $p = 0.5$ rather than at either extreme. At $\eta = 1$, the gaps are 0.575 nats ($p = 0.3$), 0.871 nats ($p = 0.5$), and 0.610 nats ($p = 0.7$). This pattern follows directly from the binomial degree distribution $\text{Bin}(n-1, p)$, whose variance $(n-1)p(1-p)$ is maximised at $p = 0.5$, producing the most heterogeneous degree profile among the three configurations. The non-monotone pattern is direct evidence that the degree proxy tracks the underlying leverage asymmetry rather than mean degree, providing empirical grounding for the degree proxy of Section V-E.

The relative-gap perspective complements the absolute picture. The largest absolute reduction in the privacy bound at $\eta = 1$ occurs on star and BA topologies (≈ 1.2 nats), where uniform DP-SGD places clients under a catastrophic bound. The largest fractional improvement occurs on ER at $p = 0.5$ (19.4% reduction), where moderate degree heterogeneity produces a leverage spread that FULCRUM exploits efficiently without the K_{uniform} baseline becoming so large as to dilute the relative gain. Together, these two perspectives show that FULCRUM delivers the largest absolute gains where the privacy risk is most severe, and the largest fractional gains where leverage asymmetry is moderate but meaningful.

D. Privacy–Utility Pareto Dominance

Having confirmed the theoretical predictions of Theorem V.3 on the synthetic η-sweep, we now evaluate whether the strict dominance of topology-aware allocation transfers to real cross-silo deployments across the full range of utility budgets and observation windows. Figure 4 shows the privacy bound $K^\star$ as a function of utility budget U across three observation windows $T_{\max}$ for each setting. Topology-aware allocation lies strictly below uniform DP-SGD at every tested U and $T_{\max}$ cell across all three settings, with no exception.

Setting A produces the largest gaps in absolute terms, with the privacy bound reduced by up to 1.967 nats (21.7%) at the strong-privacy regime ($T_{\max} = 100$, $U = 0.05$). The near 30:1 site-size heterogeneity of Fed-ISIC2019 gives the dataset-size proxy its widest leverage spread across the three settings. Setting B delivers gaps of up to 0.628 nats in absolute terms and 14.9% in relative terms, with the largest relative reduction at $T_{\max} = 50$, $U = 0.4$. Setting C, evaluated on the imbalanced hierarchical topology at $\eta = 0.5$, produces gaps up to 0.422 nats and 10.8%, reflecting the half-strength coupling; the full coupling ceiling for this topology (0.456 nats at $\eta = 1$) is confirmed by the η-sweep of Section VI-C.

A natural question is whether these privacy improvements come at the cost of model utility. Table II reports the paired utility comparison across all three settings, using TOST equivalence testing at a ± 0.5 percentage-point margin as the primary statistic. Across all three settings and all nine $T_{\max}$ rows, every TOST p-value is below 0.05, establishing statistical

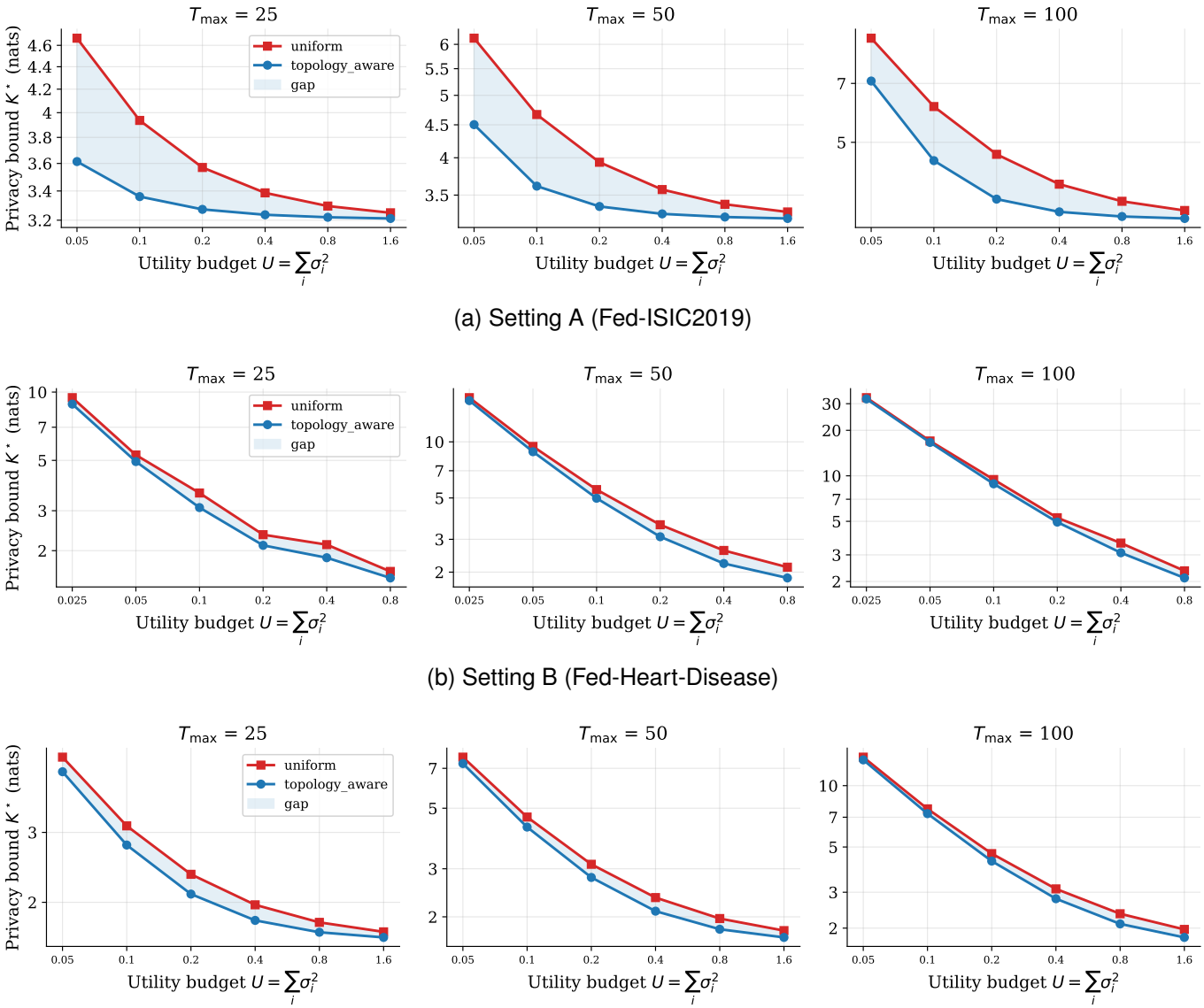

Fig. 4. Privacy–utility Pareto curves across all three settings. Each panel within a subfigure fixes $T_{\max}$; the x-axis is the utility budget U and the y-axis is the privacy bound $K^\star$ (nats, log scale). Topology-aware allocation (blue) lies strictly below uniform DP-SGD (red) at every U and $T_{\max}$ cell across all settings. The shaded region marks the absolute gap $K_{\text{uniform}} - K^\star$.

equivalence within the ± 0.5 percentage-point margin. The mean absolute accuracy difference $\overline{|\Delta|}$ does not exceed 0.486 pp on any row and collapses to 0.006 pp on Setting B at $T_{\max} = 100$. The privacy improvements are therefore obtained at no practically meaningful utility cost.

E. TADI *Channel Decomposition and Bound Realisability*

With the defense results established, we turn to the channel decomposition evaluation. The four-channel TADI decomposition serves as the empirical instantiation of Theorem V.2's additive structure, separating the controllable mechanism term from the uncontrollable prior-coupling term. The parameter channel $\mathcal{A}_1$ corresponds to the controllable term; the organisational and combined channels carry the prior-coupling term. Figure 5 compares channel lift and AUROC across the four channels on Settings B and C, spanning the two extremes of prior alignment: mismatched prior on Setting B and matched prior on Setting C at the $\eta = 1$ ceiling. Figure 6 then traces the full transition between these extremes by tracking channel lift and AUROC as a function of η on Setting C.

Three findings map directly onto the theoretical decomposition of Theorem V.2. The parameter channel $\mathcal{A}_1$ achieves negative channel lift in every configuration: -0.058 on Setting B and -0.021 on Setting C at $\eta = 1$, remaining negative across the full η range from -0.017 at $\eta = 0$ to -0.021 at $\eta = 1$. The parameter-only regressor cannot improve on the constant-mean baseline regardless of setting, prior regime, or coupling strength, confirming that the controllable mechanism term of Theorem V.2 is effectively bounded by the per-client noise allocation under both FULCRUM and uniform DP-SGD. The structural channel $\mathcal{A}_2^{\text{topo}}$ achieves near-zero lift and chance-level AUROC across both settings: in Setting C,

TABLE II

UTILITY CONSISTENCY ACROSS ALL THREE SETTINGS. EACH ROW AGGREGATES PAIRED (U, seed) OBSERVATIONS AT FIXED $T_{\max}$. $\overline{|\Delta|}$ IS THE MEAN ABSOLUTE ACCURACY DIFFERENCE IN PERCENTAGE POINTS BETWEEN TOPOLOGY-AWARE AND UNIFORM ALLOCATION AT MATCHED (U, seed). TOST p TESTS H_0: $|\mathbb{E}[\Delta]| \geq 0.5$ PP; VALUES BELOW 0.05 ESTABLISH STATISTICAL EQUIVALENCE WITHIN THE MARGIN. THE CONVENTIONAL PAIRED-t p IS REPORTED FOR COMPLETENESS.

Setting	$T_{\max}$	$\overline{\lvert\Delta\rvert}$ (pp)	95% CI (pp)	paired-t p	TOST p	n
A (Fed-ISIC2019)	25	0.399	$[-0.324, +0.322]$	0.993	0.002	18
	50	0.437	$[-0.437, +0.106]$	0.215	0.009	18
	100	0.486	$[-0.372, +0.272]$	0.745	0.005	18
B (Fed-Heart-Disease)	25	0.163	$[-0.125, +0.236]$	0.534	1.4×10^{-5}	27
	50	0.016	$[-0.017, +0.049]$	0.331	1.0×10^{-16}	18
	100	0.006	$[-0.003, +0.015]$	0.168	8.7×10^{-37}	27
C (Synthetic CIFAR-10)	25	0.016	$[-0.019, +0.009]$	0.484	2.8×10^{-23}	18
	50	0.012	$[-0.011, +0.007]$	0.614	1.6×10^{-26}	18
	100	0.016	$[-0.013, +0.010]$	0.740	8.5×10^{-25}	18

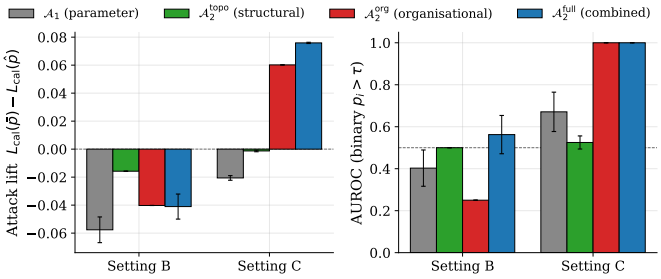

Fig. 5. Mean channel lift (left) and AUROC (right) per channel per setting, with 95% confidence intervals. Setting C bars correspond to the $\eta = 1$ matched-prior ceiling; Setting B uses the native FLamby partitioning throughout. The dashed line marks zero lift and chance AUROC ($= 0.5$) respectively. DP-SGD bounds the parameter channel $\mathcal{A}_1$ across both settings. Under matched prior, the organisational and combined channels achieve positive lift and perfect AUROC; under mismatched prior, no channel achieves positive lift.

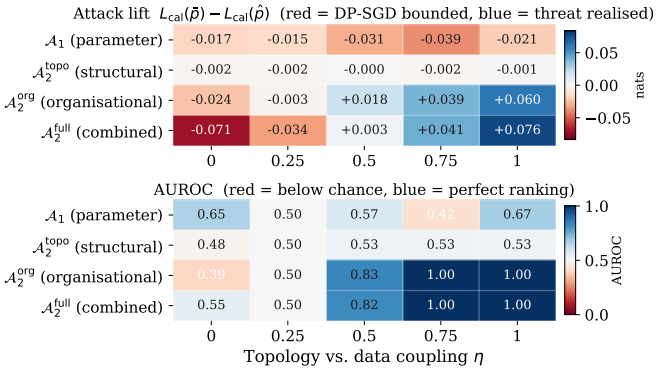

Fig. 6. Channel lift (top) and AUROC (bottom) across the (channel, η) grid on Setting C. Red cells indicate DP-SGD-bounded or below-chance performance; blue cells indicate realised leakage. The parameter channel $\mathcal{A}_1$ row is uniformly red in the lift panel at every η. The organisational and combined channels progress from red to blue as η grows, reaching lift $+0.060$ and $+0.076$ and perfect AUROC at $\eta = 1$. The $\eta = 0$ column is the IID-null calibration band; the structural channel $\mathcal{A}_2^{\text{topo}}$ row is uniformly near-white in both panels.

the hierarchical topology routes the class signal through the organisational label rather than structural graph features; in Setting B, the ring topology renders all structural features degenerate. The structural channel therefore contributes no marginal information in either deployment.

The prior-coupling channels $\mathcal{A}_2^{\text{org}}$ and $\mathcal{A}_2^{\text{full}}$ reveal the deployment-dependence of the uncontrollable term. On Setting B, where the shadow corpus is built from synthetic Dirichlet re-partitioning of public data while the target uses native FLamby site labels, both channels achieve negative lift (-0.040 and -0.041 respectively). The shadow prior does not match the target prior, and the supremum ℓ_i° is not realised. This is not a failure of the bound: it demonstrates that ℓ_i° is conservative in a deployment-favourable direction when the adversary lacks access to the native partition structure. On Setting C, the organisational channel grows monotonically in η, from -0.024 at $\eta = 0$ to $+0.060$ at $\eta = 1$, with AUROC reaching perfect ranking at $\eta \geq 0.75$. At the matched-prior ceiling of $\eta = 1$, the combined channel achieves lift $+0.076$ and perfect AUROC of 1.00. The regressor can perfectly rank all 50 clients by sensitive-class concentration, confirming that the supremum ℓ_i° is realisable when the shadow prior matches the deployment. This monotone progression is the empirical signature of the uncontrollable prior-coupling term: as the deployment encodes more class information into structural position, the lateral information $I(p_i; D_{-i})$ that DP-SGD cannot suppress grows accordingly. At $\eta = 0$, every channel's lift is slightly negative, in the range $[-0.071, -0.002]$, confirming that the regressors extract no signal when no topology-data correlation exists.

The four channels are best understood as a decomposition of information sources rather than a tournament of strictly ordered adversaries: which channel carries signal is determined by how the deployment routes the topology-data correlation, not by any universal ordering. Together, the cross-setting comparison and the η-sweep establish the three empirical questions posed at the close of Section IV. DP-SGD effectively bounds the parameter channel independently of prior coupling. The organisational label carries the prior-coupling signal in the hierarchical Setting C deployment, while neither channel carries signal in the ring Setting B. The prior-coupling supremum ℓ_i° is realisable under matched-prior conditions and conservative in a deployment-favourable direction under the public-proxy threat model of real cross-silo deployments.

VII. Discussion

Corollary V.4 establishes that FULCRUM strictly improves on uniform DP-SGD whenever leverage scores are non-uniform, and degenerates to equality when they are not. The experimental results of Section VI confirm this characterisation across the full topology spectrum. Symmetric configurations produce zero gap by construction, while every asymmetric configuration produces a strictly positive and monotonically growing gap whose magnitude scales with the degree of leverage heterogeneity. The dataset-size proxy delivers the largest absolute gaps on the real cross-silo settings, up to 1.967 nats on Setting A and 0.628 nats on Setting B, driven by the near 30:1 and 6.6:1 site-size ratios respectively. In every case the privacy gain is obtained at no practically meaningful utility cost. FULCRUM can therefore be adopted unconditionally: it provides a strict improvement wherever structural asymmetry exists and degenerates gracefully to the uniform baseline where it does not.

The empirical behaviour of the prior-coupling term warrants closer examination. Theorem V.2 defines ℓ_i° as a supremum over the prior family $\mathcal{F}_{\mathcal{G},\omega}$, representing the worst-case coupling between a client's sensitive-class concentration and the rest of the federation's data. Achieving this supremum empirically requires the shadow training distribution to match the target's deployment prior. Section VI-E provides direct evidence on this point. On Setting C, where shadow and target priors match by construction, the prior-coupling supremum is realised: the organisational channel achieves perfect AUROC and positive lift that scales monotonically with η, vanishing at $\eta = 0$ as the IID-null calibration requires. On Settings A and B, where the shadow corpus uses only synthetic Dirichlet re-partitioning of data from the same domain while the target uses native FLamby site labels, no channel achieves positive lift despite three rounds of methodological strengthening. The supremum is not realised because the shadow prior does not match the deployment's. This is a deployment-favourable property of the bound, not a weakness: the theoretical guarantee holds as an upper bound regardless of prior alignment, while the empirical result demonstrates that the worst case is not realisable by the most realistic public-proxy construction.

The prior-mismatch result raises a natural question about the validity of the shadow training setup itself. The setup assumes the adversary has access to data drawn from the same distribution as the federation without knowing the target's per-client partitioning. This is the canonical same-distribution shadow scenario of Shokri et al. [7] and is realistic in cross-silo settings where an infrastructure observer or external party may have domain knowledge, such as access to publicly released aggregate statistics, de-identified datasets from the same clinical domain, or data from a separate federation operating on similar populations, without knowledge of how any specific federation partitions that data across its participating sites. The per-client partitioning, which determines which patients are enrolled at which site, is the private quantity the adversary seeks to infer about. The IID-null condition provides the key empirical safeguard: if the regressor were exploiting global distributional properties rather than genuine partition-level signal, channel lift would be positive at $\eta = 0$ where the partitioning carries no topology-data correlation. The negative lift at $\eta = 0$ across every channel confirms that the regressors learn partition structure rather than distributional identity.

Beyond the noise allocation itself, FULCRUM is compatible with secure aggregation [28] as a complementary defense. Secure aggregation masks individual client updates so that only their aggregate is visible to the server, collapsing TADI's parameter channel to neighbourhood-aggregated observations. This weakens the controllable mechanism term of Theorem V.2 without affecting the prior-coupling term, since the structural and organisational channels depend on $(\mathcal{G}, \omega)$ rather than on Θ. FULCRUM's per-client noise allocation is performed before aggregation and is therefore fully compatible with secure aggregation, with the two mechanisms addressing different threat axes. A full empirical evaluation of the combined defense is left to future work.

Three limitations bound the scope of the claims. First, the bound of Theorem V.2 is conservative: it drops the non-negative term $I(p_i; D_{-i} \mid \Theta) \geq 0$ in the chain-rule decomposition and uses the Cuff-Yu max-KL conversion [26], which is loose at extreme noise levels. Under the Poisson subsampling applied by Opacus [35], amplified RDP [25] would tighten the per-round contribution from $1/(2\sigma_i^2|B|^2)$ to $1/(2\sigma_i^2|\mathcal{D}_i|^2)$, a substantial improvement when $|\mathcal{D}_i| \gg |B|$. Second, the optimality guarantee of Theorem V.3 and the dominance result of Corollary V.4 are stated in terms of the abstract leverage scores $\{\ell_i^\circ\}$. In practice, FULCRUM uses one of the three proxies of Section V-E in place of ℓ_i°. The guarantee is preserved in full when the proxy upper-bounds the true leverage for every client, since the allocation then injects at least as much noise as the true optimum. The group-size proxy satisfies this condition asymptotically under the stochastic block model prior. The dataset-size proxy is ordering-consistent under the Dirichlet-FedAvg generative model, which ensures Corollary V.4 applies, but whether it upper-bounds ℓ_i° in general is unproven. The degree proxy is in the same position. In deployments where a proxy underestimates ℓ_i° for some client, the per-client bound of Theorem V.2 for that client is not guaranteed; the Pareto dominance results of Section VI-D then constitute empirical rather than theoretical evidence of improvement. Third, the analysis assumes static topologies and disjoint client datasets. Dynamic reconfiguration and cross-device deployments where the same individual contributes to multiple clients both require strengthened conditions outside the current framework. The threat model is also restricted to a passive, honest-but-curious adversary; active adversaries, malicious aggregators, and Sybil attacks are complementary threats for which FULCRUM is compatible with, but not sufficient for, a complete defense.

VIII. Conclusion

This paper has shown that the communication topology of a federated learning system is an information channel that existing privacy mechanisms do not address. We introduced TADI, a shadow-trained channel decomposition with four ablations that isolates topology-conditional leakage into parameter, structural, and organisational components, and FULCRUM,

a closed-form balanced min-max optimal noise allocation derived from a per-client mutual-information bound separating a controllable mechanism term from an uncontrollable prior-coupling floor. The allocation strictly dominates uniform DP-SGD whenever the federation's leverage profile is asymmetric, degenerates exactly to uniform DP-SGD when it is not, and can therefore be adopted unconditionally. Across three benchmarks and six topology families, FULCRUM delivers privacy gains of up to 1.967 nats at no measurable utility cost. The TADI channel decomposition confirms the additive structure of the bound: the parameter channel is bounded by DP-SGD across all settings, the prior-coupling channel is empirically attained under matched-prior conditions, and the bound is conservative in a deployment-favourable direction on real cross-silo deployments.

Extending the framework to dynamic topologies, evaluating the combined defense under secure aggregation, and integrating structural leverage with per-user privacy preferences are natural next steps toward a complete treatment of topology-conditional privacy in federated learning.

DATA AND CODE AVAILABILITY

Code to reproduce this work is available at https://doi.org/10.5281/zenodo.20507155.

REFERENCES

[1] M. J. Sheller, B. Edwards *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, p. 12598, July 2020.

[2] I. Dayan, H. R. Roth *et al.*, "Federated learning for predicting clinical outcomes in patients with covid-19," *Nature medicine*, vol. 27, no. 10, pp. 1735–1743, 2021.

[3] W. Heyndrickx, L. Mervin *et al.*, "Melloddy: cross-pharma federated learning at unprecedented scale unlocks benefits in qsar without compromising proprietary information," *Journal of chemical information and modeling*, vol. 64, no. 7, pp. 2331–2344, 2023.

[4] K. Bonawitz, H. Eichner *et al.*, "Towards federated learning at scale: System design," *Proceedings of machine learning and systems*, vol. 1, pp. 374–388, 2019.

[5] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in neural information processing systems*, vol. 32, 2019.

[6] J. Geiping, H. Bauermeister *et al.*, "Inverting gradients-how easy is it to break privacy in federated learning?" *Advances in neural information processing systems*, vol. 33, pp. 16 937–16 947, 2020.

[7] R. Shokri, M. Stronati *et al.*, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[8] L. Melis, C. Song *et al.*, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 691–706.

[9] M. Abadi, A. Chu *et al.*, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[10] H. B. McMahan, D. Ramage *et al.*, "Learning differentially private recurrent language models," in *International Conference on Learning Representations*, 2018.

[11] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 2017, pp. 263–275.

[12] L. Liu, J. Zhang *et al.*, "Client-edge-cloud hierarchical federated learning," in *ICC 2020-2020 IEEE international conference on communications (ICC)*. IEEE, 2020, pp. 1–6.

[13] M. S. H. Abad, E. Ozfatura *et al.*, "Hierarchical federated learning across heterogeneous cellular networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8866–8870.

[14] A. G. Roy, S. Siddiqui *et al.*, "Braintorrent: A peer-to-peer environment for decentralized federated learning," *ArXiv*, vol. abs/1905.06731, 2019.

[15] I. Hegedűs, G. Danner, and M. Jelasity, "Decentralized learning works: An empirical comparison of gossip learning and federated learning," *Journal of Parallel and Distributed Computing*, vol. 148, pp. 109–124, 2021.

[16] J. Beilharz, B. Pfitzner *et al.*, "Implicit model specialization through dag-based decentralized federated learning," in *Proceedings of the 22nd International Middleware Conference*, 2021, pp. 310–322.

[17] J. Wang, Z. Charles *et al.*, "A field guide to federated optimization," *arXiv preprint arXiv:2107.06917*, 2021.

[18] J. Ogier du Terrail, S.-S. Ayed *et al.*, "Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5315–5334, 2022.

[19] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Technical Report*, 2009.

[20] B. Zhao, K. R. Mopuri, and H. Bilen, "idlg: Improved deep leakage from gradients," *arXiv preprint arXiv:2001.02610*, 2020.

[21] H. Yin, A. Mallya *et al.*, "See through gradients: Image batch recovery via gradinversion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 337–16 346.

[22] L. H. Fowl, J. Geiping *et al.*, "Robbing the fed: Directly obtaining private data in federated learning with modified models," in *International Conference on Learning Representations*, 2022.

[23] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753.

[24] K. Wei, J. Li *et al.*, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE transactions on information forensics and security*, vol. 15, pp. 3454–3469, 2020.

[25] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan, "Subsampled rényi differential privacy and analytical moments accountant," in *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019, pp. 1226–1235.

[26] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 43–54.

[27] S. Asoodeh, J. Liao *et al.*, "Three variants of differential privacy: Lossless conversion and applications," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 208–222, 2021.

[28] K. Bonawitz, V. Ivanov *et al.*, "Practical secure aggregation for privacy-preserving machine learning," in *2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.

[29] Z. Jorgensen, T. Yu, and G. Cormode, "Conservative or liberal? personalized differential privacy," in *2015 IEEE 31St international conference on data engineering*. IEEE, 2015, pp. 1023–1034.

[30] S. Caldas, S. M. K. Duddu *et al.*, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.

[31] C. Briggs, Z. Fan *et al.*, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *2020 international joint conference on neural networks*. IEEE, 2020, pp. 1–9.

[32] E. Cyffers and A. Bellet, "Privacy amplification by decentralization," in *International conference on artificial intelligence and statistics*. PMLR, 2022, pp. 5334–5353.

[33] C. Feng, Y. Gao *et al.*, "From models to network topologies: A topology inference attack in decentralized federated learning," *arXiv preprint arXiv:2501.03119*, 2025.

[34] H. Hu, Z. Salcic *et al.*, "Source inference attacks in federated learning," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 1102–1107.

[35] A. Yousefpour, I. Shilov *et al.*, "Opacus: User-friendly differential privacy library in pytorch," in *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

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

[37] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[38] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[39] M. Rangwala, R. O. Sinnott, and R. Buyya, "Evidential trust-aware model personalization in decentralized federated learning for wearable iot," in *2026 IEEE 26th International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. IEEE, 2026, pp. 510–519.

[40] D. Lakens, "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses," *Social psychological and personality science*, vol. 8, no. 4, pp. 355–362, 2017.