FEDERATED LEARNING WITH ENERGY-BASED STRUCTURED PROBABILISTIC INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) enables multiple clients to collaboratively train machine learning models without sharing their local data. However, heterogeneity in client data and the presence of unreliable or adversarial updates make effective server aggregation challenging, motivating methods that adapt client contributions according to their reliability and relationships. In this work, we formulate server aggregation in FL as structured probabilistic inference using conditional random fields. We represent clients as nodes in a graph and define unary and pairwise energy potentials. Unary potentials capture individual client properties, while pairwise potentials encode similarities between clients. This formulation treats individual client evidence and pairwise relationships as coupled evidence in a joint inference problem, allowing client reliabilities to be inferred collectively and the resulting estimates to guide aggregation. Where most FL methods target a single challenge, FedCRF's inferred reliabilities and affinities provide one common representation for global aggregation, Byzantine robustness, clustered FL, and personalized FL. With a single set of hyperparameters, it is competitive with or outperforms specialized methods on CIFAR-10, FEMNIST, and Fed-ISIC2019 across diverse forms of client heterogeneity and unreliability.

1 INTRODUCTION

Federated learning (FL) enables multiple clients to collaboratively train machine learning models under the coordination of a central server while keeping their data local (McMahan et al., 2017; Kairouz et al., 2021). A standard approach is Federated Averaging (FedAvg), where clients train locally from a shared model and the server aggregates the resulting models in proportion to their number of training examples (McMahan et al., 2017). This procedure provides a simple foundation for learning from decentralized data, but sample counts alone do not determine how client contributions should be combined. Differences in data distributions, data quality, and local optimization can produce conflicting updates, while malicious clients may deliberately submit harmful ones. The same observed disagreement can therefore arise from different causes and call for different aggregation strategies. Consequently, effective aggregation requires the server to account for both the reliability of individual contributions and the relationships between them.

Existing FL methods largely address these cases through specialized mechanisms. Client-side approaches constrain local optimization through proximal terms or control variates (Li et al., 2020; Karimireddy et al., 2020), while server-side aggregation methods normalize unequal local work (Wang et al., 2020) or adapt client weights (Li et al., 2023; Shi et al., 2024; Wang et al., 2023). Robust aggregation methods limit the influence of malicious updates using robust statistics or update selection (Blanchard et al., 2017; Yin et al., 2018; Pillutla et al., 2022). When a single global model cannot adequately represent all clients, clustered FL learns models for groups of related clients (Sattler et al., 2021; Ghosh et al., 2020; Ruan & Joe-Wong, 2022; Guo et al., 2024), and personalized FL adapts models to individual clients (Zhang et al., 2021). These approaches typically introduce separate mechanisms for aggregation, robustness, clustering, and personalization, although each relies on assessing which client contributions should be trusted or combined. Across these settings, existing methods rely on individual client scores or inter-client relationships as direct signals, but do not explicitly model how the two forms of evidence jointly determine client reliability.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

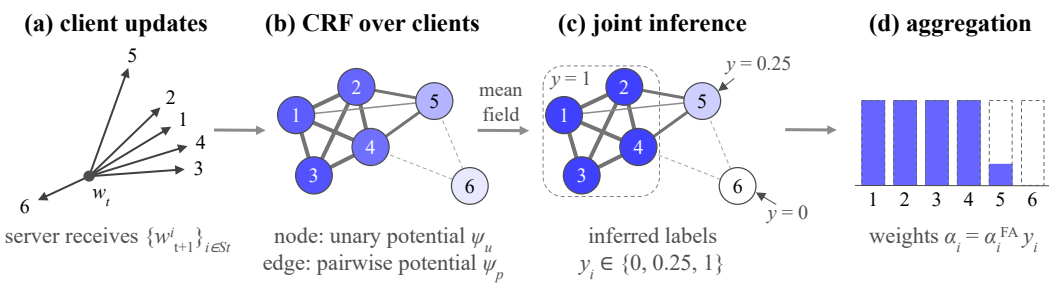


Figure 1: Illustration of the proposed method. (a) Client updates provide evidence for (b) a CRF with unary reliability potentials and pairwise client affinities. (c) Joint inference assigns labels that retain, down-weight, or exclude contributions. (d) These labels rescale the FedAvg weights.

In this work, we formulate the server's aggregation decision as *structured probabilistic inference* using a Conditional Random Field (CRF) (Lafferty et al., 2001), which we call FedCRF. We represent the participating clients as nodes in a fully connected graph and associate each client with a latent reliability label controlling whether its contribution should be retained, down-weighted, or suppressed. We define a client CRF energy consisting of *unary* and *pairwise* potentials. Unary potentials encode evidence about individual clients, while pairwise potentials encode affinities between them. Rather than using individual client scores or pairwise similarities directly, FedCRF treats them as evidence for jointly inferring these latent reliabilities over this graph, so that evidence about one client can affect the assessment of related clients. The resulting posterior reliabilities are then used to correct the standard FedAvg aggregation weights.

We instantiate this formulation using two complementary sources of evidence. *Geometric potentials* require only the returned client models and updates: individual update statistics provide unary evidence and update similarities provide pairwise affinities. *Functional potentials* instead use a small server-side proxy dataset: changes in proxy loss provide unary evidence, while agreement between client predictions defines pairwise affinity. The two sources can also be combined.

Crucially, the inferred client reliabilities and affinities provide a common representation that can be reused across different FL aggregation settings. For global aggregation and Byzantine robustness, inferred reliabilities adjust the influence of client contributions. For clustered FL, the affinity structure is used to identify groups of related clients, within which reliability-aware aggregation is performed. For personalized FL, affinities and reliabilities jointly determine client-specific model mixtures. FedCRF therefore provides a shared inference layer for interpreting client contributions, rather than requiring a separate client-assessment mechanism for each aggregation setting.

Our contributions are threefold. **First**, we formulate server-side aggregation as structured probabilistic inference over a client graph, jointly inferring latent client reliabilities from individual evidence and pairwise relationships. **Second**, we introduce geometric and functional constructions of the CRF potentials and show that the same inferred reliability and affinity representation supports *global aggregation, Byzantine robustness, clustered FL, and personalized FL* with a single set of hyperparameters. To the best of our knowledge, FedCRF is the first method to address all four settings through one shared client-assessment mechanism, whereas prior work targets one or two of them with dedicated designs. **Third**, we evaluate FedCRF on *CIFAR-10, FEMNIST, and Fed-ISIC2019* under distribution shifts, heterogeneous client quality, partial participation, malicious updates, and group- and client-level heterogeneity, comparing against specialized methods for the corresponding settings. Our implementation, including the configuration of every experiment, is available at `https://anonymous.4open.science/r/fedcrf-code-0044`.

2 Background

2.1 Conditional Random Fields

Conditional random fields (CRFs) are probabilistic graphical models that define a conditional distribution over structured data given observations (Lafferty et al., 2001). Let $G = (V, E)$ be a graph with $|V| = N$ nodes and edges E. For each node $i \in V$, we observe an input $x_i \in \mathcal{X}$ and associate

a random variable $y_i \in \mathcal{Y}$. A CRF defines the conditional distribution

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\big(-E(\boldsymbol{y} \mid \boldsymbol{x})\big), \quad E(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_i \psi_u(y_i, x_i) + \sum_{i<j} \psi_p(y_i, y_j, x_i, x_j), \quad (1)$$

where $E(\boldsymbol{y} \mid \boldsymbol{x})$ is the energy of a configuration, and $Z(\boldsymbol{x})$ is the normalizing partition function.

The *unary potential* $\psi_u(y_i, x_i)$ measures the cost of assigning label y_i to node i given its observation x_i. It is often constructed as a negative log-probability, so labels better supported by individual evidence receive lower energy. The *pairwise potential* $\psi_p(y_i, y_j, x_i, x_j)$ encodes interactions between nodes. A common parameterization is

$$\psi_p(y_i, y_j, x_i, x_j) = \mu(y_i, y_j) \sum_{m=1}^{M} a_m k_m(x_i, x_j), \quad (2)$$

where $k_m(x_i, x_j) \geq 0$ are observation-dependent affinities, $a_m \geq 0$ are their weights, and $\mu(y_i, y_j)$ specifies the compatibility of a label pair.

The *maximum a posteriori* (MAP) configuration minimizes the energy $\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y} \mid \boldsymbol{x}) = \arg\min_{\boldsymbol{y}} E(\boldsymbol{y} \mid \boldsymbol{x})$. For general potentials on dense graphs, exact MAP inference and exact marginal computation are computationally intractable. A common approach to approximate marginal inference is the *mean-field approximation* (Wainwright & Jordan, 2008; Krähenbühl & Koltun, 2011), which introduces a node-factorized distribution $q(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_i q_i(y_i \mid \boldsymbol{x})$ and minimizes $\mathrm{KL}(q|p)$.

2.2 Federated Learning

Consider K clients, where client k owns a local dataset $\mathcal{D}_k$ containing n_k examples, with $n = \sum_{k=1}^{K} n_k$ examples in total. In the standard empirical FL formulation, the goal is to collaboratively learn a shared model parameter vector $w \in \mathbb{R}^d$ across clients by minimizing

$$\min_w F(w) = \sum_{k=1}^{K} p_k F_k(w), \qquad p_k = \frac{n_k}{n}, \qquad F_k(w) = \frac{1}{n_k} \sum_{(x,y)\in\mathcal{D}_k} \ell(w; x, y). \quad (3)$$

Other client-weighted objectives are possible, but weighting clients by their number of local examples is the standard choice underlying FedAvg (McMahan et al., 2017).

FedAvg optimizes this objective through repeated rounds of local training and server aggregation. At communication round t, the server broadcasts the current model w_t to a subset of participating clients S_t. Each participating client initializes from w_t, performs local optimization on its dataset, and returns a locally trained model w_{t+1}^k. The server then aggregates the returned models as

$$w_{t+1} = \sum_{k\in S_t} \alpha_k^{\mathrm{FA}} w_{t+1}^k, \qquad \alpha_k^{\mathrm{FA}} = \frac{n_k}{\sum_{j\in S_t} n_j}. \quad (4)$$

Federated aggregation settings. FL methods differ in how client contributions are combined. *Global aggregation* learns a single model shared across clients, while *clustered FL* addresses structured heterogeneity by learning shared models for groups of related clients (Sattler et al., 2021; Ghosh et al., 2020; Ruan & Joe-Wong, 2022; Guo et al., 2024). *Personalized FL* operates at a finer granularity, allowing client-specific models while retaining collaboration across clients (T. Dinh et al., 2020; Li et al., 2021; Zhang et al., 2021). These settings mainly differ in how they accommodate client heterogeneity and the desired degree of model sharing. Orthogonally, *Byzantine-robust aggregation* considers that some clients may submit arbitrary or adversarial updates and seeks to limit their influence (Blanchard et al., 2017; Yin et al., 2018; Pillutla et al., 2022). Thus, disagreement among client contributions may motivate different strategies, such as weighting, grouping, personalization, or robust aggregation, which are typically addressed by specialized methods. We discuss these approaches in detail in Section A.

3 Method

We introduce **FedCRF**, a method that formulates FL server-side aggregation as structured probabilistic inference over client contributions. Starting from FedAvg, we retain local training and use a CRF to infer correction coefficients for the aggregation weights.

3.1 CRF FOR FEDERATED LEARNING

At each round t, we represent the participating clients as a fully connected graph $G = (V, E)$, with one node per client $i \in S_t$. Each node has a latent label $y_i \in \mathcal{Y}$ that specifies how to adjust its contribution: retain its unnormalized FedAvg weight, reduce it, or exclude it, respectively. In aggregation, we use the posterior expectation of this label as the correction coefficient (details in Section 3.3). The observation x_i contains evidence from the client's update $u_i = w_{t+1}^i - w_t$ and, when available, its model's predictions on a server-side proxy dataset. We omit the round index below for simplicity.

To instantiate the CRF, we define the unary potential, pairwise potential, and label compatibility function as

$$\psi_u(y_i, x_i) = -\log \pi_i(y_i), \quad \psi_p(y_i, y_j, x_i, x_j) = \lambda A_{ij} \mu(y_i, y_j), \quad \mu(y_i, y_j) = (y_i - y_j)^2. \tag{5}$$

Here, π_i is a prior distribution over the client's labels, $A_{ij} \geq 0$ is a symmetric affinity between clients, and $\lambda \geq 0$ controls the strength of the pairwise term. The squared label difference makes disagreement more costly for strongly connected clients and distinguishes partial down-weighting from complete exclusion.

We construct the unary prior from an individual reliability score $\rho_i \in (0, 1)$:

$$\pi_i(y) = \frac{b_i(y)}{\sum_{\ell \in \mathcal{Y}} b_i(\ell)}, \qquad b_i(y) = \rho_i^{\,y} \, (1 - \rho_i)^{\,1-y}. \tag{6}$$

The endpoint scores are $b_i(1) = \rho_i$ and $b_i(0) = 1 - \rho_i$, while $b_i(0.25) = \rho_i^{1/4}(1 - \rho_i)^{3/4}$ is between them. Thus, larger reliability scores favor retaining a contribution, and smaller scores favor excluding it, while at $\rho_i = \frac{1}{2}$ the prior is uniform over the three labels. We next specify how client evidence determines ρ_i and A_{ij}.

3.2 GEOMETRIC AND FUNCTIONAL POTENTIALS

Geometric evidence: FedCRF_G. We use client updates $u_i = w_{t+1}^i - w_t$ and their coordinate-wise median $\tilde{u} = \mathrm{coordmed}_{j \in S_t} u_j$, requiring no server-side data. For each client, we define cosine alignment, normalized distance, and norm anomaly as

$$c_i = \cos(u_i, \tilde{u}), \qquad d_i = \frac{\|u_i - \tilde{u}\|_2}{\mathrm{med}_j \, \|u_j - \tilde{u}\|_2}, \qquad h_i = \frac{\big|\|u_i\|_2 - \mathrm{med}_j \, \|u_j\|_2\big|}{\mathrm{MAD}_j(\|u_j\|_2)}. \tag{7}$$

Here, $\mathrm{MAD}_j(z_j) = \mathrm{med}_j \, |z_j - \mathrm{med}_k \, z_k|$, all medians are over participating clients. The reliability and affinities are

$$\rho_i^G = \sigma(2c_i - d_i - 0.5h_i), \qquad A_{ij}^G = \exp\left(-\frac{1 - \cos(u_i, u_j)}{\theta_G}\right) \min(\rho_i^G, \rho_j^G), \tag{8}$$

where σ is the logistic sigmoid, $\theta_G > 0$, and $A_{ii}^G = 0$. The affinity combines directional agreement with a reliability gate that limits strong connections within unreliable clients and between reliable and unreliable ones.

Functional evidence: FedCRF_F. This variant assumes a small proxy dataset $\mathcal{D}_{\mathrm{prox}}$ at the server and constructs evidence from model performance and predictive agreement. For each client, we measure the improvement in proxy loss,

$$\delta_i = \mathcal{L}_{\mathrm{prox}}(w_t) - \mathcal{L}_{\mathrm{prox}}(w_{t+1}^i), \qquad \rho_i^F = \sigma(\hat{\delta}_i / \tau), \tag{9}$$

where $\hat{\delta}_i$ is standardized using the median and median absolute deviation across participating clients, and $\tau > 0$ is a temperature.

Let $p_i(\cdot \mid x)$ denote the predictive distribution of w_{t+1}^i. We compute the empirical mean symmetric divergence on proxy inputs and the corresponding affinity:

$$D_{ij} = \frac{1}{2} \mathbb{E}_{x \sim \mathcal{D}_{\mathrm{prox}}} \left[\mathrm{KL}(p_i \| p_j) + \mathrm{KL}(p_j \| p_i) \right], \qquad A_{ij}^F = \exp(-D_{ij}/\theta_F) \min(\rho_i^F, \rho_j^F), \tag{10}$$

with zero diagonal. Thus, unary evidence measures relative improvement in proxy performance, while pairwise evidence measures agreement between predictions.

Mixed evidence: FedCRF_{G+F}**.** We also combine the two constructions:

$$\rho_i^{G+F} = \frac{\rho_i^G + \rho_i^F}{2}, \qquad A_{ij}^{G+F} = \sqrt{A_{ij}^G A_{ij}^F}, \quad i \neq j. \tag{11}$$

3.3 Inference to aggregation

We approximate the conditional distribution using the mean-field updates from Section 2.1. Starting from $q_i^{(0)}(y) = \pi_i(y)$, we perform R mean-field updates:

$$q_i^{(r+1)}(y) = \frac{\exp\Big[-\psi_u(y, x_i) - \lambda \sum_{j \neq i} A_{ij} \sum_{\ell \in \mathcal{Y}} q_j^{(r)}(\ell)(y - \ell)^2\Big]}{\sum_{z \in \mathcal{Y}} \exp\Big[-\psi_u(z, x_i) - \lambda \sum_{j \neq i} A_{ij} \sum_{\ell \in \mathcal{Y}} q_j^{(r)}(\ell)(z - \ell)^2\Big]}. \tag{12}$$

The posterior expectation $\bar{y}_i$ provides the correction coefficient for client i. Using the FedAvg weights from Equation (4), the server computes

$$\bar{y}_i = \sum_{y \in \mathcal{Y}} y\, q_i^{(R)}(y), \qquad \alpha_i = \frac{\alpha_i^{\mathrm{FA}} \bar{y}_i}{\sum_{j \in S_t} \alpha_j^{\mathrm{FA}} \bar{y}_j}, \qquad w_{t+1} = \sum_{i \in S_t} \alpha_i w_{t+1}^i. \tag{13}$$

FedCRF therefore adjusts sample-count weights using jointly inferred reliabilities. Equal positive coefficients recover FedAvg.

3.4 Extensions to Group-Wise and Client-Specific Aggregation

Clustered FL. To adapt for Clustered FL setting, we use client similarities before reliability gating to get locally scaled affinities $\tilde{A}_{ij}$ (Zelnik-manor & Perona, 2004). Then we use spectral clustering to identify client groups, with their number specified or selected by the eigengap. Each group then trains a shared model using FedCRF aggregation among its members.

Personalized FL. When clients do not form clear groups, the goal can instead be to learn a separate model for each client. The server still uses pairwise affinities to share model weights between similar clients, but each client receives its own combination of the returned models. After local training from personalized models w_i^{pers}, the server updates the distance graph, recomputes the locally scaled affinities, and mixes the returned models as

$$w_i^{\mathrm{pers}} \leftarrow \sum_j W_{ij} w_{t+1}^j, \qquad W_{ij} = \frac{B_{ij} \bar{y}_j}{\sum_\ell B_{i\ell} \bar{y}_\ell}, \tag{14}$$

where $B_{ij} = \tilde{A}_{ij}^{1/T}$ for $j \neq i$ and $B_{ii} = s$. Temperature $T > 0$ controls neighbor selectivity, and $s > 0$ controls self-weight. Each client's mixing weights therefore reflect both its affinities to other clients and their inferred reliability.

4 Experiments

4.1 Setup

Datasets and models. We evaluate controlled heterogeneity on CIFAR-10 (ResNet-9) and natural federation on FEMNIST and Fed-ISIC2019. ANDA (Li et al., 2025) partitions CIFAR-10 by label skew $P(y)$, feature skew $P(x)$, concept shift $P(y \mid x)$, or class-conditional feature shift $P(x \mid y)$, at medium and high severity, alongside independent and identically distributed (IID) data. Its *strict* variants provide known client groups for evaluating clustering. FEMNIST (Caldas et al., 2019) uses writers as clients (62 classes, LEAF CNN). Fed-ISIC2019 from FLamby (du Terrail et al., 2022) uses an ImageNet-pretrained EfficientNet-B0 for 8 classes; its 6 acquisition centres are split into sub-clients, with centre identity as the reference grouping.

Training setup. We implement FedCRF and all baselines in PyTorch in a single code base, so that all methods share the data partitions, local training, attacks, and evaluation. Unless stated otherwise, we use 20 clients with full participation, 2 local epochs, batch size 64, and SGD with learning rate 0.01 and momentum 0.9. Training runs for at most 300 rounds, with early stopping after 5 rounds without improvement in sample-weighted client validation loss. Each client's data is split 80/10/10 for training, validation, and testing; results use 3 seeds. The proxy set contains 10 examples per class from the official test split, excluded from evaluation. Our main metric is sample-weighted client test accuracy; we additionally report official test accuracy on the remaining benchmark test set. Default hyperparameter values are $(\lambda, \theta_G, \theta_F, \tau) = (0.5, 0.5, 1, 1)$, $\mathcal{Y} = \{0, 0.25, 1\}$, $R = 5$, and personalization self-weight $s = 10$. We use the same set of hyperparameters for all the datasets and settings. FedCRF is not sensitive to hyperparameters choice and we did not tune them for any of the tables: across the 21 settings of the sensitivity grid, validation accuracy varies by at most 1.6 points (see Appendix Section F). Tables mark the best mean performance in bold and the second best underlined. Extra details are available in Appendix Section B, we also include all the required to reproduce the results in our anonymized repo.

Baselines. We compare global aggregation with FedAvg, FedProx (Li et al., 2020), FedNova (Wang et al., 2020), FedLAW (Li et al., 2023), FedLWS (Shi et al., 2024), and FedCDA (Wang et al., 2023). Robustness baselines include coordinate-wise median, trimmed mean (Yin et al., 2018), Krum and Multi-Krum (Blanchard et al., 2017), RFA (Pillutla et al., 2022), geometric median, FedNGA (Zuo et al., 2025), FLTrust (Cao et al., 2021), and Zeno (Xie et al., 2019); the latter two use the same proxy set as FedCRF. For clustering and personalization, we include local training, FedAvg with local fine-tuning, Ditto (Li et al., 2021), pFedGraph (Ye et al., 2023), IFCA (Ghosh et al., 2020), CFL (Sattler et al., 2021), FedSoft (Ruan & Joe-Wong, 2022), FedRC (Guo et al., 2024), FedFomo (Zhang et al., 2021), and FLUX (Fenoglio et al., 2025). A more detailed description of the baselines is given in Section B.5.

4.2 Global aggregation

Heterogeneous clients. Table 1 reports result for federated training with heterogenous clients. For CIFAR-10, we use ANDA (Li et al., 2025) to generate non-IID client distributions, while FEMNIST and ISIC2019 provide naturally heterogeneous data. In case of no data corruption (see *clean* columns in Table 1), all three versions of the proposed approach generally perform on par with the best baselines, and even outperform them slightly in the most complex case of ISIC2019.

To test the methods in a more difficult setup, we corrupt the training data of 30% of clients by injecting standard Gaussian noise into the input images, leaving validation and test data clean (see *corrupted* in Table 1). Across the six clean and corrupted settings, FedCRF_G and FedCRF_{G+F} have the highest mean accuracy (73.6% and 73.0%, versus 71.9% for the best baseline). The clearest gains occur on corrupted CIFAR-10 and Fed-ISIC2019, where all three variants outperform all the baselines, led by FedCRF_F on CIFAR-10 at 82.7% versus 78.1% for FedLWS. On Fed-ISIC2019, FedCRF_G gets the best result and reaches 62.0% versus 53.5% for FedLWS, while functional and mixed evidence offer smaller gains but still outperform the best baseline. We further show in Appendix Section H how each evidence mode suppresses corrupted clients.

Partial participation. Table 2 shows results under partial client participation, where at each round only a small random cohort participates. Our method with functional and mixed evidence achieves the highest averages across the six settings in Table 2: 65.5% and 65.2%, versus 63.5% for FedProx. On CIFAR-10 $P(y)$ medium with 50 clients and 20% participation, FedCRF_F reaches 71.6%, compared with 54.1% for FedAvg and 70.2% for FedLAW, the strongest baseline in this case. This advantage is not uniform: with 100 clients and 10% participation, FedLAW leads at 61.2%, while all other methods remain below 33%. However, this strong performance does not transfer consistently across datasets, as FedLAW drops to 5.2% and 6.2% on FEMNIST and to 57.6% and 54.2% on Fed-ISIC2019 under 20% and 10% participation, respectively, while the other methods remain around 75–78% and 67–68.5%. On Fed-ISIC2019, whose six centres we split into 50 or 100 sub-clients, FedCRF_F and FedCRF_{G+F} are the most accurate with 100 clients (67.9%), and FedCRF_{G+F} is within 0.2 points of the best method with 50 clients (68.3% versus 68.5% for FedProx).

Table 1: Global aggregation under heterogeneous clients. In *corrupted* case, 30% of the clients train on data corrupted with Gaussian noise. Client test accuracy (%, ↑), 3 seeds.

Method	CIFAR-10, $P(y)$ medium		FEMNIST (20 writers)		Fed-ISIC2019		
	clean	corrupted	clean	corrupted	clean	corrupted	avg
Baselines							
FedAvg	84.5±0.3	77.5±0.8	**74.0±2.9**	72.4±4.6	68.3±0.7	53.3±0.9	71.7
FedProx	84.7±0.7	77.3±1.1	74.0±2.7	72.7±5.1	68.3±0.4	53.3±1.0	71.7
FedNova	84.8±0.4	77.4±1.1	73.6±1.6	72.1±5.3	67.9±1.0	53.2±0.9	71.5
FedLWS	**85.3±0.9**	78.1±0.9	74.0±2.8	72.5±5.1	68.2±1.1	53.5±0.9	71.9
FedLAW	77.4±6.8	72.7±0.6	5.0±1.1	5.6±0.8	49.1±0.8	49.1±0.8	43.1
Ours							
FedCRF_G	84.6±0.9	80.8±0.7	73.3±4.5	71.4±5.1	**69.7±0.9**	**62.0±3.4**	**73.6**
FedCRF_F	84.3±1.2	**82.7±1.0**	71.7±5.3	71.1±3.4	69.2±0.5	54.4±3.4	72.2
FedCRF_{G+F}	84.7±0.5	82.0±1.8	73.7±4.3	**72.9±3.3**	69.1±0.5	55.5±2.4	73.0

Table 2: Partial participation: client test accuracy (%, ↑) with K clients and the given participation rate per round (10-client cohorts), 3 seeds. Fed-ISIC2019 splits its six centres into K sub-clients in proportion to their size.

Method	CIFAR-10, $P(y)$ medium		FEMNIST		Fed-ISIC2019		
	K=50, 20%	K=100, 10%	K=50, 20%	K=100, 10%	K=50, 20%	K=100, 10%	avg
Baselines							
FedAvg	54.1±10.8	31.8±4.7	77.1±3.1	**78.2±1.6**	68.1±2.2	67.4±1.1	62.8
FedProx	64.1±18.6	25.6±11.2	77.8±2.9	78.0±1.4	**68.5±1.8**	66.8±1.7	63.5
FedNova	63.4±16.7	18.7±2.5	75.5±1.5	76.7±1.1	67.7±1.8	67.3±1.4	61.6
FedLWS	68.1±4.4	17.0±5.2	**77.8±2.7**	78.0±1.5	67.6±1.7	66.9±1.0	62.6
FedLAW	70.2±9.4	**61.2±3.0**	5.2±0.7	6.2±1.0	57.6±1.1	54.2±5.9	42.4
Ours							
FedCRF_G	57.4±14.1	24.7±7.9	75.8±1.6	77.9±1.1	68.1±1.3	67.6±0.6	61.9
FedCRF_F	**71.6±3.5**	32.9±2.2	75.3±3.2	77.5±1.4	67.9±1.2	**67.9±0.4**	**65.5**
FedCRF_{G+F}	71.0±14.2	30.9±3.2	75.7±2.3	77.2±0.8	68.3±1.3	67.9±0.7	65.2

4.3 Byzantine robustness

In Table 3 we evaluate our method under five attacks with 20% fixed malicious clients: sign flipping, ALIE (Baruch et al., 2019), targeted label flipping, model replacement (Bagdasaryan et al., 2020), and Neurotoxin (Zhang et al., 2022). We report clean accuracy and, for the targeted attacks, attack success rate (ASR).

Functional evidence provides the strongest overall results, while other variants still perform better than most of the baselines on average (except for Zeno which gets second best score). FedCRF_F averages 83.4% clean accuracy across five attacks, on par with Zeno's 83.1% and above the best baseline without server data (79.9%; Table 3). Under ALIE, FedCRF_F reaches 83.6%, essentially matching Zeno at 83.7% with the same proxy set, compared with 76.7% for geometric evidence and 70.6% for trimmed mean.

Under model replacement and Neurotoxin, FedCRF_F combines 83.3–83.4% clean accuracy with ASR of 2.5% and 1.9%. Coordinate-wise median, trimmed mean, RFA, and geometric median retain high clean accuracy but exceed 89% ASR. Overall, FedCRF achieves a strong balance between high accuracy and low ASR. Reliability diagnostics further confirm that malicious clients are assigned substantially lower inferred reliability over training rounds (Appendix H).

4.4 Clustered and Personalized FL

Group-structured heterogeneity. In these experiments we group clients in a way, that within a group the clients have IID data, and non-IID compared to clients from other groups. We consider two settings with and without ground-truth number of groups J provided, our method supports

Table 3: Byzantine robustness on CIFAR-10 ($P(y)$ medium, 20% fixed malicious clients, 3 seeds): clean accuracy (%) under five attacks and attack success rate (ASR, %) for the targeted attacks.

Method	Clean accuracy under attack ↑						Attack success rate ↓			
	sign flip	ALIE	targeted LF	model repl.	Neurotoxin	avg	targeted LF	model repl.	Neurotoxin	avg
Robust aggregation										
Coord. median	78.4±1.8	54.0±13.5	83.2±1.3	84.8±0.4	**85.0±1.0**	77.1	3.8±1.3	94.9±0.2	93.0±1.9	63.9
Trimmed mean	73.9±3.5	70.6±1.3	82.8±1.3	**85.1±0.9**	84.9±0.3	79.5	9.9±4.1	95.7±0.7	94.7±0.9	66.8
Krum	46.6±24.1	27.6±9.3	70.2±4.5	51.6±1.8	51.6±1.8	49.5	1.9±0.7	6.5±4.2	6.5±4.2	5.0
Multi-Krum	**83.9±0.7**	65.5±4.9	82.5±1.5	83.4±1.0	83.4±1.0	79.8	0.7±0.4	**1.5±0.4**	1.5±0.4	**1.3**
RFA	81.8±0.4	68.5±2.6	80.9±4.0	83.8±0.8	84.6±0.9	79.9	6.5±2.4	92.4±1.7	89.4±1.4	62.8
Geom. median	81.8±0.4	68.5±2.6	80.9±4.0	83.8±0.8	84.6±0.9	79.9	6.5±2.4	92.4±1.7	89.4±1.4	62.8
Fed-NGA	9.9±0.1	65.2±3.3	77.0±0.9	57.4±39.9	10.4±0.9	44.0	13.6±4.5	60.1±52.1	**0.0±0.0**	24.6
With server data										
FLTrust	73.1±0.1	76.4±1.4	77.4±0.8	77.7±1.4	77.6±1.4	76.5	3.4±1.0	73.4±3.0	2.3±0.6	26.4
Zeno	83.4±0.5	**83.7±0.9**	82.8±0.5	82.8±0.3	83.0±0.2	83.1	0.8±0.3	1.9±0.3	1.6±0.6	1.4
Ours										
FedCRF$_G$	81.5±0.7	76.7±1.6	82.8±1.1	82.7±0.7	83.3±0.6	81.4	4.0±2.9	32.6±27.2	8.0±3.4	14.9
FedCRF$_F$	83.0±0.6	83.6±0.7	**83.7±1.0**	83.4±0.2	83.3±1.9	**83.4**	**0.6±0.1**	2.5±0.4	1.9±0.5	1.7
FedCRF$_{G+F}$	82.0±0.3	77.9±1.3	83.5±0.4	82.9±1.9	83.4±1.0	81.9	1.0±0.2	12.5±16.1	6.2±6.1	6.6

Table 4: Clustered FL on group-structured heterogeneity: client test accuracy (%) and adjusted Rand index (ARI) between discovered and true groups, 3 seeds. Methods in the oracle-J groups receive the true number of clusters, the others estimate it.

Method	CIFAR $P(y)$ strict, med.		CIFAR $P(y)$ strict, high		FEMNIST, rotated		FEMNIST, label-swapped		Fed-ISIC2019 (centres)		average	
	acc ↑	ARI ↑	acc ↑	ARI ↑	acc ↑	ARI ↑	acc ↑	ARI ↑	acc ↑	ARI ↑	acc ↑	ARI ↑
Baselines												
FedAvg	46.5±4.8	–	35.6±19.4	–	60.8±3.6	–	35.0±1.3	–	68.3±0.7	–	49.2	–
CFL	39.8±8.4	0.00	39.2±16.8	0.00	**64.9±4.3**	0.53	63.7±10.4	0.87	68.3±0.7	0.00	55.2	0.28
FLUX	81.0±3.0	0.49	92.4±5.4	0.47	49.9±1.7	0.02	62.2±2.0	0.38	**72.6±1.6**	**0.83**	71.6	0.44
Ours												
FedCRF$_G$	38.1±5.9	0.00	46.8±10.7	0.25	55.9±2.4	0.02	36.1±2.7	0.00	72.4±1.3	0.58	49.9	0.17
FedCRF$_F$	59.9±33.6	0.44	63.9±14.3	0.47	44.3±27.8	0.04	27.7±7.3	0.00	69.2±0.8	0.00	53.0	0.19
FedCRF$_{G+F}$	84.8±5.0	0.95	60.7±18.4	0.24	57.8±3.6	0.02	33.8±3.3	0.00	71.7±1.4	0.58	61.8	0.36
Baselines with oracle J												
IFCA	76.9±6.2	0.80	77.1±11.1	0.67	62.0±4.2	0.33	58.5±2.7	0.67	57.3±1.5	0.00	66.4	0.49
FedSoft	74.2±3.5	0.68	86.1±12.9	0.60	60.9±1.4	**0.95**	61.4±3.8	**1.00**	65.9±1.6	0.44	69.7	**0.73**
FedRC	72.3±5.0	0.75	84.8±7.8	0.93	62.3±0.6	0.00	40.9±10.6	0.15	59.1±1.6	0.22	63.9	0.41
FedFomo	63.4±7.2	–	88.4±7.1	–	41.7±0.5	–	45.6±7.2	–	69.4±0.7	–	61.7	–
FLUX	75.7±8.5	0.63	90.7±10.0	0.88	56.5±2.3	0.02	56.7±4.5	0.50	70.1±1.3	0.48	69.9	0.50
Ours with oracle J												
FedCRF$_G$	**87.5±2.2**	**1.00**	95.6±2.6	**1.00**	48.5±4.9	0.07	64.1±5.9	**1.00**	71.3±0.4	0.44	**73.4**	0.70
FedCRF$_F$	57.9±32.5	0.62	94.7±2.6	**1.00**	23.6±16.3	0.02	53.7±7.1	0.89	64.9±3.2	0.27	59.0	0.56
FedCRF$_{G+F}$	58.6±33.1	0.47	**95.7±3.0**	**1.00**	60.3±2.4	0.11	**64.7±5.4**	**1.00**	71.1±1.0	0.47	70.1	0.61

both. Table 4 compares accuracy and Adjusted Rand Index (ARI) on strict CIFAR-10 label skew, grouped FEMNIST, and Fed-ISIC2019. With the true group count, FedCRF$_G$ has the highest average accuracy of all methods across the five settings (73.4%, versus 71.6% for the strongest baseline, FLUX). It recovers the CIFAR-10 groups exactly (ARI 1.00), reaching 87.5% and 95.6% at medium and high severity, versus 76.9% and 90.7% for the strongest baselines given the same count. It also recovers FEMNIST label-permutation groups exactly, but fails on rotation groups (ARI 0.07, accuracy 48.5%). With the true count, mixed evidence is the most accurate method under high-severity CIFAR-10 label skew (95.7%) and on label-swapped FEMNIST (64.7%), and improves on geometric evidence for rotated FEMNIST (60.3%) even without accurate definition of those groups. Without the true count, FedCRF$_{G+F}$ is the only method that recovers the medium-severity CIFAR-10 groups (ARI 0.95) and the most accurate one there (84.8%, versus 81.0% for FLUX). On Fed-ISIC2019, FedCRF$_G$ finds two groups and reaches 72.4%, on par with FLUX (72.6%).

Per-client heterogeneity. We consider no group-structure among clients. The goal is to train best possible per-client models while still benefiting from collaboration across clients. FedCRF uses the affinities for client-specific mixing (Table 5). FedCRF with mixed evidence achieves the highest four-setting average (74.6%), followed by functional evidence (74.1%) and pFedGraph (72.7%). The clearest gain occurs under class-conditional feature shift $P(x|y)$, where mixed evidence reaches 84.8%, compared with 81.8% for pFedGraph. Functional evidence is also the most accurate under label skew, reaching 93.6%, slightly above local training (93.4%) and FLUX (93.3%).

Table 5: Personalized FL on per-client heterogeneity (CIFAR-10 at ANDA high severity; FEMNIST with 20 writers): client test accuracy (%, ↑), 3 seeds.

Method	ANDA high severity (per-client shift)				
	CIFAR $P(y)$	CIFAR $P(y \mid x)$	CIFAR $P(x \mid y)$	FEMNIST	avg
Baselines					
FedAvg	14.2±3.0	52.5±2.3	55.4±12.1	74.0±2.9	49.0
Local only	93.4±2.7	39.5±2.9	74.9±3.3	56.9±2.4	66.2
FedAvg + FT	91.4±2.1	**58.4±2.4**	74.3±8.8	62.8±3.9	71.7
Ditto	91.6±1.8	45.7±1.7	68.6±22.9	67.7±4.5	68.4
pFedGraph	86.3±1.7	51.9±3.4	81.8±3.9	70.6±2.7	72.7
IFCA	20.8±0.6	53.0±2.7	68.9±10.7	74.3±3.2	54.2
CFL	12.9±1.4	53.7±1.2	62.1±8.6	72.1±3.4	50.2
FedSoft	90.7±2.6	41.9±2.2	57.6±16.9	61.2±4.1	62.8
FedRC	18.4±7.6	53.3±1.7	79.3±1.2	**76.0±2.7**	56.8
FedFomo	90.5±2.5	22.4±15.4	60.5±4.9	48.7±1.7	55.5
FLUX	93.3±1.9	44.4±2.9	75.2±3.0	62.3±3.4	68.8
Ours					
FedCRF$_G$	78.2±9.3	51.0±2.5	83.6±1.7	66.4±1.2	69.8
FedCRF$_F$	**93.6±2.3**	51.9±2.4	81.3±1.8	69.5±0.7	74.1
FedCRF$_{G+F}$	90.6±1.2	53.5±2.2	**84.8±1.4**	69.4±0.2	**74.6**

5 Discussion and limitations

FedCRF provides a common inference framework for client reliability and relationships that can be reused across global aggregation, Byzantine robustness, clustered FL, and personalized FL. While on average FedCRF consistently outperforms baselines, some specialized approaches can edge it in specific cases.

While FedCRF is effective against the attacks considered on CIFAR-10, its robustness does not transfer uniformly across datasets. In particular, the FEMNIST results (see Appendix Table 6) show that some backdoor attacks can retain high attack success despite strong clean accuracy, a failure mode also observed for many robust aggregation baselines. This suggests a broader difficulty in detecting localized or trigger-dependent behavior from coarse whole-model update statistics or clean proxy evaluation. A promising extension is therefore to enrich both sources of evidence: geometric evidence could be constructed at a finer granularity, for example layer-wise or over parameter groups, so that localized anomalies are less likely to be diluted in the overall model signal. Functional evidence could actively probe returned models using perturbed or augmented proxy inputs and measure their prediction stability or disagreement, potentially exposing behaviors that remain hidden on clean inputs. In our work we also use a fixed set of hyperparameters, which potentially can be made trainable for even better versatility.

Clustered and personalized cases depend on the affinity representation and group-count estimate, both of which may fail in some settings. The CRF on a dense graph introduces computational overhead over standard FedAvg. Appendix G shows, however, that this overhead is comparable to several lightweight server-side baselines and lower than many specialized alternatives.

6 Conclusion

FedCRF formulates server-side aggregation as structured probabilistic inference over client contributions, jointly modeling individual reliability and inter-client relationships from geometric or functional evidence. The resulting reliabilities and affinities provide a common representation that can be reused across global aggregation, Byzantine robustness, clustered FL, and personalized FL. Across heterogeneous client quality, partial participation, malicious updates, and structured or client-specific heterogeneity, FedCRF achieves competitive or leading performance against specialized methods. These findings support structured reliability modeling as a flexible foundation for federated aggregation across diverse FL settings.

AI USE STATEMENT

We used large language model (LLM) assistants in three ways: to help implement the method and the baselines, to help set up, run, and report the experiments (experiment configurations, job orchestration, and the scripts that build the result tables), and to proofread and edit the text. The research questions, the method, the experimental design, and the interpretation of the results are our own. We take full responsibility for the content of the paper

REPRODUCIBILITY STATEMENT

The code, the batch configuration of every experiment, and the result workbooks behind every table are available at `https://anonymous.4open.science/r/fedcrf-code-0044`. The repository contains one script per table and for the diagnostics figure, each of which reruns the corresponding experiments and regenerates the table, and a script that regenerates all tables and the figure from the released results without training.

REFERENCES

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How To Backdoor Federated Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948. PMLR, June 2020.

Gilad Baruch, Moran Baruch, and Yoav Goldberg. A Little Is Enough: Circumventing Defenses For Distributed Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A Benchmark for Federated Settings, December 2019.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. In *Network and Distributed System Security Symposium (NDSS)*, 2021.

Fengwen Chen, Guodong Long, Zonghan Wu, Tianyi Zhou, and Jing Jiang. Personalized Federated Learning With a Graph. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2575–2582. International Joint Conferences on Artificial Intelligence Organization, 2022. ISBN 978-1-956792-00-3. doi: 10.24963/ijcai.2022/357. URL `https://www.ijcai.org/proceedings/2022/357`.

Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, et al. FLamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 5315–5334, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Dario Fenoglio, Mohan Li, Pietro Barbiero, Nicholas Lane, Marc Langheinrich, and Martin Gjoreski. FLUX: Efficient Descriptor-Driven Clustered Federated Learning under Arbitrary Distribution Shifts. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 71229–71292. Curran Associates, Inc., 2025. doi: 10.52202/085713-2395.

Dario Fenoglio, Martin Gjoreski, Edoardo Gabrielli, and Marc Langheinrich. Security and Privacy in Federated Learning: A Survey and Unified Taxonomy of Threat Models, Attacks, and Defenses. *Social Science Research Network (SSRN)*, 2026. doi: 10.2139/ssrn.6771005.

Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. The Limitations of Federated Learning in Sybil Settings. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*. USENIX Association, 2020.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An Efficient Framework for Clustered Federated Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19586–19597. Curran Associates, Inc., 2020.

Yongxin Guo, Xiaoying Tang, and Tao Lin. FedRC: Tackling Diverse Distribution Shifts Challenge in Federated Learning by Robust Clustering. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 16910–16944. PMLR, July 2024.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, et al. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 14(1-2):1–210, June 2021. ISSN 1935-8237. doi: 10.1561/2200000083.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5132–5143. PMLR, November 2020.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, pp. 109–117, Red Hook, NY, USA, December 2011. Curran Associates Inc. ISBN 978-1-61839-599-3.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 282–289, San Francisco, CA, USA, June 2001. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-778-1.

Mohan Li, Dario Fenoglio, Martin Gjoreski, and Marc Langheinrich. Federated Learning with Profile Mapping under Distribution Shifts and Drifts. In *The Fourteenth International Conference on Learning Representations*, October 2025.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. *Proceedings of Machine Learning and Systems (MLSys)*, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and Robust Federated Learning Through Personalization. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6357–6368. PMLR, July 2021.

Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. Revisiting Weighted Aggregation in Federated Learning with Neural Networks. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19767–19788. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23s.html.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, April 2017.

Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, et al. FLAME: Taming Backdoors in Federated Learning. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1415–1432, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-31-1. URL https://www.usenix.org/conference/usenixsecurity22/presentation/nguyen.

Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust Aggregation for Federated Learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022. doi: 10.1109/TSP.2022.3153135.

Yichen Ruan and Carlee Joe-Wong. FedSoft: Soft Clustered Federated Learning with Proximal Local Updating. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8124–8131, 2022.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3710–3722, August 2021. ISSN 2162-2388. doi: 10.1109/TNNLS.2020.3015958.

Changlong Shi, Jinmeng Li, He Zhao, Dan dan Guo, and Yi Chang. FedLWS: Federated Learning with Adaptive Layer-wise Weight Shrinking. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Canh T. Dinh, Nguyen Tran, and Josh Nguyen. Personalized Federated Learning with Moreau Envelopes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21394–21405. Curran Associates, Inc., 2020.

Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, January 2008. ISSN 1935-8237. doi: 10.1561/2200000001.

Haozhao Wang, Haoran Xu, Yichen Li, Yuan Xu, Ruixuan Li, and Tianwei Zhang. FedCDA: Federated Learning with Cross-rounds Divergence-aware Aggregation. In *The Twelfth International Conference on Learning Representations*, October 2023.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7611–7623. Curran Associates, Inc., 2020.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed Stochastic Gradient Descent with Suspicion-based Fault-tolerance. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6893–6901. PMLR, May 2019.

Rui Ye, Zhenyang Ni, Fangzhao Wu, Siheng Chen, and Yanfeng Wang. Personalized Federated Learning with Inferred Collaboration Graphs. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 39801–39817. PMLR, July 2023.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5650–5659. PMLR, July 2018.

Lihi Zelnik-manor and Pietro Perona. Self-Tuning Spectral Clustering. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M. Alvarez. Personalized Federated Learning with First Order Model Optimization. In *International Conference on Learning Representations*, October 2021.

Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable Backdoors in Federated Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 26429–26446. PMLR, June 2022.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional Random Fields as Recurrent Neural Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1529–1537, December 2015. doi: 10.1109/ICCV.2015.179.

Shiyuan Zuo, Xingrun Yan, Rongfei Fan, Li Shen, Puning Zhao, Jie Xu, and Han Hu. Efficient Federated Learning against Byzantine Attacks and Data Heterogeneity via Aggregating Normalized Gradients. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.

APPENDIX

A RELATED WORK

Server-side weighting. FedNova normalizes unequal local work (Wang et al., 2020), FedLAW learns aggregation weights and a global step size on proxy data (Li et al., 2023), FedLWS applies layer-wise shrinkage from update statistics (Shi et al., 2024), and FedCDA selects cached client models by divergence (Wang et al., 2023). FedCRF instead treats aggregation weights as outputs of inference over latent client reliabilities, combining individual scores with pairwise compatibility.

Robust aggregation. Robust aggregation is one family within the broader landscape of FL defenses against adversarial behavior (Fenoglio et al., 2026). Coordinate-wise median, trimmed mean (Yin et al., 2018), Krum and Multi-Krum (Blanchard et al., 2017), and RFA (Pillutla et al., 2022) use robust statistics or update selection. Zeno scores updates using validation loss (Xie et al., 2019), while FLTrust compares them with an update from a server-held root dataset (Cao et al., 2021). Our functional potentials similarly use server data and add predictive agreement between clients; on CIFAR-10, functional FedCRF matches the strongest server-data defense, Zeno, while also providing the client affinities used for clustering and personalization. Similarity also underlies FoolsGold's detection of coordinated clients (Fung et al., 2020) and FLAME's update clustering (Nguyen et al., 2022). FedCRF uses similarity to couple reliability labels, with a gate that limits reinforcement between individually unreliable clients.

Clustered and personalized FL. Clustered methods learn group models through client assignments or memberships (IFCA, FedSoft, FedRC (Ghosh et al., 2020; Ruan & Joe-Wong, 2022; Guo et al., 2024)), update-similarity partitions (CFL (Sattler et al., 2021)), or data descriptors (FLUX (Fenoglio et al., 2025)). Personalized methods regularize client models toward a shared model (Ditto, pFedMe (Li et al., 2021; T. Dinh et al., 2020)) or mix models using client-specific weights (FedFomo (Zhang et al., 2021)). Graph-based approaches such as SFL and pFedGraph explicitly model client relationships (Chen et al., 2022; Ye et al., 2023). FedCRF reuses the affinities constructed for reliability inference to support either clustering or personalized mixing, with inferred reliabilities adjusting client contributions.

CRFs and mean-field inference. Dense CRFs combine individual evidence with pairwise label compatibility through approximate inference (Krähenbühl & Koltun, 2011), which can also be unrolled as recurrent layers (Zheng et al., 2015). FedCRF applies this formulation to clients, using latent reliability labels to determine their aggregation weights.

B EXPERIMENTAL SETUP

B.1 DATASETS AND MODELS

CIFAR-10 and MNIST. CIFAR-10 contains 50,000 training and 10,000 test RGB images of size 32×32 in 10 classes. We use ResNet-9 and ANDA partitions (Section B.2). The ablations in Sections E and F also use MNIST with the same partitioning scheme and an MLP with hidden layers of size 256 and 128.

FEMNIST. We use LEAF's Federated EMNIST (Caldas et al., 2019): 28×28 grayscale images in 62 classes, grouped by writer. We retain writers with at least 100 examples, reserve a random 20% of writers for the official test set, and sample clients from the rest ($K = 20$ by default). The LEAF CNN has two 5×5 convolutional layers with 32 and 64 channels, each followed by max pooling, and a 2048-unit fully connected layer before the classifier.

Grouped FEMNIST. For Table 4, we randomly assign 20 writers to four groups of five. Each group applies either a rotation of $g \cdot 90^\circ$ or a fixed permutation of the 62 labels, with group $g = 0$ using the identity permutation. The transformation applies to all client splits; proxy and official test data remain unchanged. Group assignments provide the reference labels for ARI.

Fed-ISIC2019. This FLamby benchmark (du Terrail et al., 2022) contains 8 skin-lesion classes from 6 acquisition centres. We use FLamby's train/test split and resizing, followed by a 200×200 centre crop, and an ImageNet-pretrained EfficientNet-B0 with a new classifier. Centres are divided into 18 sub-clients in proportion to their size, with at least one per centre. Centre identity defines the reference groups. We also report balanced accuracy because of class imbalance.

Splits, proxy set, and metrics. Client data is split 80/10/10 for training, validation, and testing. Early stopping uses sample-weighted client validation loss. Client test accuracy is sample-weighted across clients, each evaluated with its assigned global, cluster, or personalized model. FedCRF's functional variants, FedLAW, FLTrust, and Zeno share a proxy set of 10 examples per class from the official test split; these examples are excluded from official test evaluation.

B.2 Heterogeneity with ANDA

ANDA (Li et al., 2025) varies one factor of $P(x, y)$ across clients at a time. We use static partitions; paired values below specify medium and high severity.

- *Label skew* $P(y)$: clients receive equal shares of the remaining data pool, sampled with class probabilities $\mathrm{softmax}(n_c^s)$, where n_c is the remaining class count and s is sampled uniformly from $[0.5, 1]$ / $[1, 3]$ for each client.
- *Feature skew* $P(x)$: client-specific distributions over rotations and colours, using 2 rotations and 2 colours / 4 rotations and 3 colours.
- *Concept shift* $P(y \mid x)$: each client permutes labels within a randomly chosen pool of 3 / 5 classes.
- *Class-conditional feature shift* $P(x \mid y)$: each client rotates and recolours 3 / 5 randomly chosen classes, using 2 rotations and 2 colours / 4 rotations and 3 colours.

Default partitions assign each client its own shift. The *strict* variants instead use known groups: strict label skew assigns clients to three label subsets of 4 / 2 classes, while strict feature skew assigns a rotation–colour combination from 2 rotations and 2 colours / 4 rotations and 3 colours.

B.3 Client quality and participation

For Table 1 and the diagnostics in Section H, we corrupt only the training split of 30% of randomly chosen clients. Label noise replaces 50% or 80% of labels with a uniformly sampled different class; input noise adds Gaussian noise with standard deviation 1 to normalized inputs. For Table 2, each round samples pK clients uniformly. Sampling depends only on the seed and round, ensuring identical cohorts across methods.

B.4 Attacks

We select 20% of clients as malicious once per run; they attack from the first round.

- *Sign flipping*: submit $w_t - 2u_i$ instead of $w_t + u_i$.
- *ALIE* (Baruch et al., 2019): submit the coordinate-wise mean of honest updates minus z times their standard deviation, with z determined by the participating and malicious client counts.
- *Targeted label flipping*: change class 1 to class 0. ASR is the fraction of class-1 test examples predicted as class 0.
- *Model replacement* (Bagdasaryan et al., 2020): use the same backdoor training and scale the submitted update by 10.
- *Neurotoxin* (Zhang et al., 2022): use the same backdoor training and restrict the update to coordinates outside the largest 5% of the previous global update in absolute magnitude.

B.5 BASELINES

All methods use the training setup in Section 4.1, with the following aggregation or local-objective changes.

- *Global aggregation.* FedAvg (McMahan et al., 2017) uses sample-count weights. FedProx (Li et al., 2020) adds $\frac{\mu}{2}\|w - w_t\|^2$ to the local objective ($\mu = 0.01$). FedNova (Wang et al., 2020) normalizes by local step count. FedLAW (Li et al., 2023) learns weights and global shrinkage on the proxy set using 100 server epochs of Adam. FedLWS (Shi et al., 2024) applies layer-wise shrinkage ($\beta = 0.03$). FedCDA (Wang et al., 2023) selects the least-divergent combination from three cached models per client after 50 warm-up rounds.
- *Robust aggregation.* We use coordinate-wise median and trimmed mean (Yin et al., 2018), trimming 10% from each side. Krum and Multi-Krum (Blanchard et al., 2017) assume $f = \lfloor 0.2|S_t| \rfloor$ attackers and select one or $|S_t| - f - 2$ updates by nearest-neighbour distance scores. RFA (Pillutla et al., 2022) computes a sample-weighted geometric median with Weiszfeld iterations; the geometric-median baseline uses uniform weights. Fed-NGA averages unit-normalized updates.
- *Robust aggregation with server data.* FLTrust (Cao et al., 2021) trains a root update u_0 using the median client step count, scores clients by $\mathrm{ReLU}(\cos(u_i, u_0))$, rescales their updates to $\|u_0\|$, and averages with these scores. Zeno (Xie et al., 2019) scores proxy-loss decrease minus $\rho\|u_i\|^2$ ($\rho = 5 \cdot 10^{-4}$) and averages the best $|S_t| - f$ updates.
- *Clustered FL.* IFCA (Ghosh et al., 2020) assigns clients by local loss; CFL (Sattler et al., 2021) recursively partitions update similarities when training stalls; FedSoft (Ruan & Joe-Wong, 2022) uses soft memberships; FedRC (Guo et al., 2024) models distribution shifts; and FLUX (Fenoglio et al., 2025) clusters data descriptors. IFCA, FedSoft, FedRC, and oracle-count FLUX receive the true group count where defined, and two clusters otherwise. Adaptive FLUX estimates the count.
- *Personalized FL.* Local training uses no communication. FedAvg with fine-tuning evaluates each client's locally updated global model. Ditto (Li et al., 2021) regularizes a personal model toward the FedAvg model ($\lambda = 0.1$). pFedGraph (Ye et al., 2023) uses update similarity to infer a collaboration graph (prior $\alpha = 0.8$) and a cosine regularizer toward the graph-weighted model ($\lambda = 0.01$). FedFomo (Zhang et al., 2021) weights candidate models by local validation improvement, using the true group count as its candidate count where groups are defined.

B.6 FEDCRF SETTINGS

Unless varied in an ablation, we use the geometric coefficients $(2, 1, 0.5)$, $\lambda = 0.5$, $\theta_G = 0.5$, $\theta_F = \tau = 1$, $\mathcal{Y} = \{0, 0.25, 1\}$, and $R = 5$. Mixed evidence uses equal weights; functional evidence uses the proxy set above.

Clustering. We run five global warm-up rounds with all clients and no reliability gate, averaging $-\log A_{ij}$ into distances d_{ij}. Local scaling uses each client's seventh-nearest-neighbour distance (Zelnik-manor & Perona, 2004). Spectral clustering uses the normalized Laplacian, with the group count selected by the largest eigengap among the first 12 eigenvalues on CIFAR-10 and grouped FEMNIST, or 8 on Fed-ISIC2019. Oracle-count variants use the true count. Each group then trains with FedCRF aggregation.

Personalization. After the same warm-up, clients maintain individual models. Each round updates the ungated distances by an exponential moving average with rate 0.3, recomputes locally scaled affinities, and mixes models using Equation (14), with all clients as neighbours, $T = 1$, and $s = 10$.

C ALGORITHM

Algorithm 1 separates one round into local training, evidence construction, inference, and aggregation. CLIENTUPDATE performs the configured local training; evidence formulas are given in Section 3.2. All client distributions in iteration $r+1$ use the distributions from iteration r. In computation, reliability scores are clipped to $[\epsilon, 1-\epsilon]$, scale denominators are bounded below by ϵ, and label normalization uses a numerically stable softmax. Structure-discovery settings are in Section B.6.

Algorithm 1 FedCRF: one communication round.

Require: Global model w_t; clients S_t with sample counts n_i
Require: Evidence mode $m \in \{G, F, G{+}F\}$; proxy set $\mathcal{D}_{\text{prox}}$ if $m \neq G$; parameters $\mathcal{Y}, \lambda, \theta_G, \theta_F, \tau, R, \epsilon$
Ensure: Global model w_{t+1}, client reliabilities $\bar{y}$, and affinity matrix A

1. Train clients and collect updates
for all $i \in S_t$ **in parallel do**
  $w_{t+1}^i \leftarrow \text{CLIENTUPDATE}(i, w_t)$
end for
$u_i \leftarrow w_{t+1}^i - w_t, \quad \alpha_i^{\text{FA}} \leftarrow n_i / \sum_{j \in S_t} n_j$ for each client i
2. Construct evidence on the server
if $m = G$ **then**
  Compute $(\rho, A) = (\rho^G, A^G)$ from the updates using Equation (8)
else if $m = F$ **then**
  Compute $(\rho, A) = (\rho^F, A^F)$ on $\mathcal{D}_{\text{prox}}$ using Equations (9) and (10)
else
  Compute both (ρ^G, A^G) and (ρ^F, A^F) using the same formulas
  $\rho_i \leftarrow (\rho_i^G + \rho_i^F)/2, \quad A_{ij} \leftarrow \sqrt{A_{ij}^G A_{ij}^F}$ for $i \neq j$
end if
Set $A_{ii} \leftarrow 0$ for each client i
3. Infer client reliabilities
Initialize $q_i^{(0)}(y) \leftarrow \pi_i(y)$ from Equation (6) for every i and $y \in \mathcal{Y}$
for $r = 0$ **to** $R-1$ **do**
  for all $i \in S_t$ **in parallel do**
    $C_i^{(r)}(y) \leftarrow \lambda \sum_{j \neq i} A_{ij} \sum_{\ell \in \mathcal{Y}} q_j^{(r)}(\ell)(y-\ell)^2$ for each $y \in \mathcal{Y}$
    $\tilde{q}_i(y) \leftarrow \pi_i(y) \exp\big(-C_i^{(r)}(y)\big)$ for each $y \in \mathcal{Y}$
    $q_i^{(r+1)}(y) \leftarrow \tilde{q}_i(y) / \sum_{z \in \mathcal{Y}} \tilde{q}_i(z)$ for each $y \in \mathcal{Y}$
  end for
end for
$\bar{y}_i \leftarrow \sum_{y \in \mathcal{Y}} y\, q_i^{(R)}(y)$ for each client i
4. Normalize weights and aggregate
$v_i \leftarrow \alpha_i^{\text{FA}} \bar{y}_i$ for each client $i, \quad Z \leftarrow \sum_{j \in S_t} v_j$
if $Z > \epsilon$ **then**
  $\alpha_i \leftarrow v_i / Z$ for each client i
else
  $\alpha_i \leftarrow \alpha_i^{\text{FA}}$ for each client i {fall back to FedAvg}
end if
$w_{t+1} \leftarrow \sum_{i \in S_t} \alpha_i w_{t+1}^i$
return $w_{t+1}, \bar{y}, A$

D BYZANTINE ROBUSTNESS ON FEMNIST

On FEMNIST, all three FedCRF variants are among the most accurate defenses (Table 6). Mixed and functional evidence average 71.1% clean accuracy over the five attacks, within 0.5 points of Zeno (71.6%), and geometric evidence 70.6%; all three are ahead of every robust aggregator without server data (RFA 69.5%, trimmed mean 67.6%, geometric median 65.0%, Multi-Krum 62.8%).

Table 6: Byzantine robustness on FEMNIST (20 writers, 20% fixed malicious clients, 3 seeds): clean accuracy (%) under five attacks and attack success rate (ASR, %) for the targeted attacks.

Method	Clean accuracy under attack ↑						Attack success rate ↓			
	sign flip	ALIE	targeted LF	model repl.	Neurotoxin	avg	targeted LF	model repl.	Neurotoxin	avg
Robust aggregation										
Coord. median	22.9±30.8	63.9±3.6	71.7±4.8	69.6±4.9	70.6±3.8	59.7	**0.0±0.0**	95.9±4.0	95.3±5.2	63.7
Trimmed mean	47.7±8.9	67.0±4.5	**75.1±3.7**	**75.0±3.7**	73.2±4.5	67.6	2.5±2.4	99.8±0.2	99.7±0.2	67.3
Krum	12.1±7.8	12.9±16.7	4.5±1.7	14.4±14.2	6.1±0.6	10.0	**0.0±0.0**	2.0±3.5	**0.0±0.0**	**0.7**
Multi-Krum	64.0±4.3	56.5±12.6	60.4±6.7	67.0±5.4	66.0±3.5	62.8	1.6±2.7	**1.7±0.0**	13.8±7.5	5.7
RFA	57.1±5.0	**70.1±2.3**	73.5±2.9	72.9±2.9	**74.0±3.9**	69.5	1.6±2.7	99.9±0.1	99.9±0.1	67.1
Geom. median	40.5±29.8	68.3±2.8	73.0±3.0	72.2±4.6	70.9±6.6	65.0	3.5±3.1	100.0±0.0	100.0±0.0	67.8
Fed-NGA	59.8±5.6	58.2±0.6	62.3±5.8	63.5±3.5	63.4±2.7	61.5	**0.0±0.0**	99.5±0.7	99.1±1.5	66.2
With server data										
FLTrust	3.7±0.9	3.7±1.0	4.1±0.9	7.2±0.8	5.6±2.2	4.9	22.2±38.5	3.9±5.2	34.2±57.0	20.1
Zeno	**71.6±3.2**	68.7±4.5	73.2±4.7	71.3±5.1	73.4±2.7	**71.6**	1.4±2.5	98.6±1.7	100.0±0.0	66.7
Ours										
FedCRF$_G$	66.8±3.4	67.7±4.6	71.5±3.7	73.7±3.2	73.3±3.5	70.6	**0.0±0.0**	99.6±0.5	99.9±0.1	66.5
FedCRF$_F$	69.9±4.5	68.6±3.9	72.8±3.1	72.5±3.5	71.6±4.4	71.1	**0.0±0.0**	99.6±0.3	100.0±0.0	66.5
FedCRF$_{G+F}$	67.4±1.5	68.0±3.1	74.1±2.8	73.2±2.6	72.8±2.8	71.1	1.0±1.6	99.7±0.3	99.9±0.1	66.9

Table 7: Personalized FedCRF on the per-client cells (client test accuracy, %, ↑): self weight s (default s=10), neighbour count N_{nbr}, and evidence.

Method	MNIST high			CIFAR-10 high		
	$P(y)$	$P(y \mid x)$	$P(x \mid y)$	$P(y)$	$P(y \mid x)$	$P(x \mid y)$
FedCRF$_G$ (s=1)	94.7±5.6	66.8±4.1	95.4±0.9	57.0±17.5	53.0±4.8	73.9±8.1
FedCRF$_G$ (s=3)	95.8±3.0	70.9±2.9	94.0±2.2	70.4±16.0	54.2±3.0	83.1±1.5
FedCRF$_G$ (s=10)	99.5±0.6	**81.9±2.2**	94.2±0.4	78.2±9.3	51.0±2.5	83.6±1.7
FedCRF$_G$ (s=1, N_{nbr}=5)	96.5±2.8	77.6±2.3	96.1±0.3	58.1±10.7	**55.3±1.9**	75.8±2.1
FedCRF$_F$ (s=10)	**99.8±0.1**	73.9±3.4	96.1±0.4	**93.6±2.3**	51.9±2.4	81.3±1.8
FedCRF$_{G+F}$ (s=1)	97.2±2.0	65.7±2.5	96.4±0.4	78.0±8.7	54.3±3.1	55.2±31.9
FedCRF$_{G+F}$ (s=3)	99.1±0.4	67.7±3.2	96.3±0.3	81.5±14.1	53.9±3.0	78.6±3.0
FedCRF$_{G+F}$ (s=10)	99.7±0.2	78.0±3.0	**96.4±0.3**	90.6±1.2	53.5±2.2	**84.8±1.4**

FedCRF also neutralizes targeted label flipping (ASR 0.0–1.0%) while keeping 71.5–74.1% clean accuracy. The backdoors that scale or mask the malicious update (model replacement and Neurotoxin) remain effective against every aggregation rule that keeps high clean accuracy on FEMNIST, including Zeno, RFA, and trimmed mean (ASR above 95%); only Krum, Multi-Krum, and FLTrust reduce them, at a large cost in clean accuracy (10.0%, 62.8%, and 4.9% on average).

E Self weight of personalized FedCRF

Increasing self-weight improves personalized FedCRF under several shifts (Table 7). The default $s = 10$ has the highest client validation accuracy averaged over the six cells for mixed evidence (83.8%, against 80.9% for $s = 3$ and 77.3% for $s = 1$); only under concept shift do $s = 1$ and $s = 3$ validate slightly better. For mixed evidence on CIFAR-10, raising s from 1 to 3 to 10 increases accuracy from 78.0% to 81.5% to 90.6% under label skew and from 55.2% to 78.6% to 84.8% under class-conditional shift. Concept-shift accuracy changes little (53.5–54.3%). The other evidence types benefit in the same way: at $s = 10$, geometric evidence reaches 78.2% under CIFAR-10 label skew (57.0% at $s = 1$) and functional evidence 93.6%, the best result in that column, and geometric evidence is the most accurate variant under MNIST concept shift (81.9%). Restricting geometric evidence to five neighbours at $s = 1$ helps MNIST concept shift (77.6% versus 66.8%) but gives little improvement on CIFAR-10 label skew (58.1% versus 57.0%). These results motivate the default $s = 10$ with all neighbours.

Table 8: Component ablation of FedCRF_G in benign settings (client test accuracy, %, ↑). Direct ρ_i weighting sets $\alpha_i \propto \alpha_i^{\mathrm{FA}} \rho_i$, without the CRF labels or the pairwise term.

Method	CIFAR-10 $P(y)$		CIFAR-10 $P(x)$	
	med	hig	med	hig
FedAvg	84.5	<u>14.5</u>	**73.3**	<u>67.2</u>
direct ρ_i weighting	84.6±0.6	11.8±4.6	72.6±1.9	65.6±1.3
FedCRF_G without gate	**85.0±0.7**	12.0±1.2	72.2±0.5	67.1±0.7
FedCRF_G uniform base weights	84.5±1.5	13.5±4.4	72.4±1.6	**67.3±1.5**
FedCRF_G	<u>84.6±0.9</u>	**18.8±6.8**	<u>72.9±1.1</u>	67.2±1.7

F Additional ablations

Components. Table 8 compares full FedCRF_G with direct reliability weighting ($\alpha_i \propto \alpha_i^{\mathrm{FA}} \rho_i$), removal of the reliability gate, and uniform base weights. In these benign settings no client needs to be down-weighted, so all variants perform comparably; the full model is the most accurate under CIFAR-10 $P(y)$ high (18.8%). The structured components matter once some clients are unreliable: Section H shows that under ALIE the pairwise term lowers the attackers' aggregation weight in every seed and improves the worst-case accuracy (Table 13).

Inference. FedCRF is robust to its inference settings (Table 9). With $R = 1$ or $R = 10$ mean-field iterations instead of the default $R = 5$, or with two or five labels instead of three, all variants lie within 0.8 points of one another at medium severity, and at high severity on CIFAR-10 the default is the most accurate (18.8% against 10.4–13.7%).

The Gaussian relaxation uses continuous labels, $y_i \in \mathbb{R}$, a quadratic unary potential $\psi_u(y_i, x_i) = (y_i - \rho_i)^2/(2\sigma_u^2)$, and the same pairwise potential. Its posterior mean and MAP estimate coincide:

$$\bar{\boldsymbol{y}} = \left(I + 2\lambda\sigma_u^2 L\right)^{-1} \boldsymbol{\rho}, \qquad L = \mathrm{diag}(A\mathbf{1}) - A. \tag{15}$$

Here, L is the affinity-graph Laplacian; we use $\sigma_u = 1$ and clip the result to $[0, 1]$. The relaxation smooths reliability scores over the graph without iterative label inference. The two formulations are comparable at medium severity, while discrete labels are clearly more accurate under high MNIST label skew: 85.7–86.5% versus 76.8% for geometric evidence, and 85.9% versus 72.6% for mixed evidence. This motivates the discrete label set used throughout.

Hyperparameters. On CIFAR-10 $P(y)$ medium, the 21 settings of (λ, θ_G) for FedCRF_G and (θ_F, τ) for FedCRF_F span 84.0–85.6% accuracy (Table 10) and 84.8–86.4% mean client validation accuracy. This shows that sensitivity to a hyperparameter choice is modest for FedCRF.

Proxy set. FedCRF needs only a very small proxy set: on MNIST and CIFAR-10 $P(y)$ medium, one example per class performs comparably to ten, and fifty gives no further gain (Table 11). It also tolerates imperfect proxy data: label noise, class imbalance, and input noise in the proxy set reduce accuracy by at most 0.9 points relative to a clean proxy set.

Table 9: Inference variants (client test accuracy, %, ↑): mean-field iterations R, label-set size, and the Gaussian relaxation.

Method	MNIST $P(y)$		CIFAR-10 $P(y)$	
	med.	high	med.	high
FedCRF$_G$ (default: $\lvert\mathcal{Y}\rvert = 3$, $R = 5$)	97.9±0.0	85.7±4.5	84.6±0.9	**18.8±6.8**
mean field, $R = 1$	**98.0±0.1**	**86.5±4.9**	84.6±0.6	13.7±3.5
mean field, $R = 10$	97.9±0.1	86.1±5.2	85.0±0.2	10.8±3.0
mean field, $\mathcal{Y} = \{0, 1\}$	97.8±0.1	82.2±4.4	84.6±0.4	10.4±1.0
mean field, $\lvert\mathcal{Y}\rvert = 5$	97.9±0.1	82.0±6.0	**85.4±0.7**	11.0±3.7
FedCRF$_G$ (Gaussian)	97.8±0.1	76.8±6.7	84.9±0.4	11.0±0.3
FedCRF$_{G+F}$ (default)	97.9±0.1	85.9±3.6	84.7±0.5	17.5±5.9
FedCRF$_{G+F}$ (Gaussian)	97.8±0.1	72.6±5.3	84.7±0.8	12.3±1.0

Table 10: Hyperparameter sensitivity on CIFAR-10 $P(y)$ medium (client test accuracy, %).

	accuracy ↑
FedCRF$_G$: $\lambda = 0.1, \theta_G = 0.25$	84.5±1.1
FedCRF$_G$: $\lambda = 0.1, \theta_G = 0.5$	84.4±0.7
FedCRF$_G$: $\lambda = 0.1, \theta_G = 1$	84.2±1.6
FedCRF$_G$: $\lambda = 0.5, \theta_G = 0.25$	84.9±0.7
FedCRF$_G$: $\lambda = 0.5, \theta_G = 0.5$	84.6±0.9
FedCRF$_G$: $\lambda = 0.5, \theta_G = 1$	85.6±0.5
FedCRF$_G$: $\lambda = 1, \theta_G = 0.25$	84.2±0.2
FedCRF$_G$: $\lambda = 1, \theta_G = 0.5$	85.0±0.1
FedCRF$_G$: $\lambda = 1, \theta_G = 1$	84.7±0.9
FedCRF$_G$: $\lambda = 5, \theta_G = 0.25$	84.3±0.7
FedCRF$_G$: $\lambda = 5, \theta_G = 0.5$	84.5±0.9
FedCRF$_G$: $\lambda = 5, \theta_G = 1$	84.8±0.4
FedCRF$_F$: $\theta_F = 0.5, \tau = 0.5$	84.1±0.6
FedCRF$_F$: $\theta_F = 0.5, \tau = 1$	84.8±0.6
FedCRF$_F$: $\theta_F = 0.5, \tau = 2$	84.3±0.3
FedCRF$_F$: $\theta_F = 1, \tau = 0.5$	84.2±1.3
FedCRF$_F$: $\theta_F = 1, \tau = 1$	84.3±1.2
FedCRF$_F$: $\theta_F = 1, \tau = 2$	84.6±1.1
FedCRF$_F$: $\theta_F = 2, \tau = 0.5$	84.2±0.4
FedCRF$_F$: $\theta_F = 2, \tau = 1$	84.0±0.5
FedCRF$_F$: $\theta_F = 2, \tau = 2$	84.5±0.6

Table 11: Proxy-set realism ($P(y)$ medium, client test accuracy, %, ↑): proxy budget per class and proxy corruptions.

Method	MNIST		CIFAR-10	
	FedCRF$_F$	FedCRF$_{G+F}$	FedCRF$_F$	FedCRF$_{G+F}$
reference (10/class, clean proxy)	97.8±0.0	97.9±0.1	84.3±1.2	84.7±0.5
proxy budget 1/class	97.8±0.0	97.9±0.1	84.7±1.0	84.6±0.4
proxy budget 5/class	98.0±0.0	98.0±0.0	84.3±0.2	84.5±0.5
proxy budget 50/class	97.9±0.1	97.9±0.1	83.5±0.3	85.0±0.5
label noise 20%	97.9±0.1	98.0±0.1	83.5±0.7	84.5±1.1
label noise 50%	98.0±0.0	98.0±0.1	84.3±0.5	84.2±1.0
class imbalance 20%	97.8±0.1	97.9±0.1	84.4±1.1	84.8±0.8
class imbalance 50%	97.9±0.1	98.0±0.0	84.9±0.9	83.9±1.1
input noise 0.2	97.9±0.1	97.9±0.1	84.8±0.5	84.2±0.4
input noise 0.5	97.9±0.1	97.9±0.0	84.7±0.3	83.8±0.8

G COMPUTATIONAL COST

Complexity. FedCRF changes only the server step. With N participating clients and d model parameters, geometric evidence needs the coordinate-wise median of the updates ($O(Nd \log N)$), each update's distance to it ($O(Nd)$), and the pairwise cosine similarities ($O(N^2 d)$). Functional evidence evaluates every returned model on the proxy set (N forward passes over $10C$ examples for C classes) and compares the predictions of every pair of clients ($O(N^2 C^2)$). Mean-field inference on the fully connected graph costs $O(RN^2|\mathcal{Y}|^2)$, which is negligible. Like any rule that aggregates full updates, the server keeps the N updates in memory. Clustering adds one eigendecomposition of an $N \times N$ matrix after the warm-up and then aggregates within each group, while personalization forms one mixture of the N returned models per client ($O(N^2 d)$ per round).

Measurements. We ran every method for four rounds without early stopping on one NVIDIA RTX 4090 GPU (one seed, no attack, the same proxy set for all methods) in the settings of the paper: CIFAR-10 with ResNet-9 (6.6M parameters) and FEMNIST with the LEAF CNN (6.6M parameters), each with 20 and 100 clients, and Fed-ISIC2019 with EfficientNet-B0 (4.0M parameters) and 18 clients. All clients participate in every round. Table 12 reports the median wall-clock time of rounds 2–4, which includes local training, the server step, and evaluation, relative to FedAvg. FedCDA searches its model cache only after a warm-up (50 rounds in the paper); here we shortened the warm-up to one round and measured two rounds of search. Its search is exhaustive within batches of clients, so with 100 clients we used 15 batches instead of 3, which keeps about seven clients per batch as with 20 clients.

Results. The server step of FedCRF is lightweight. Global FedCRF adds 0.7–0.9 s to the server step with 20 clients on CIFAR-10 and FEMNIST (0.5–1.5 s on Fed-ISIC2019, where functional evidence evaluates EfficientNet-B0 on the proxy set) and 3.1–5.3 s with 100 clients, against 0.1 s and 0.4–0.8 s for FedAvg's own server step. How much this matters depends on the cost of local training. Per round, it is +10–13% for ResNet-9 on CIFAR-10 with 20 clients, +29–40% with 100 clients, and +1–4% for EfficientNet-B0 on Fed-ISIC2019, but +95–148% on FEMNIST, where each writer holds about 200 examples and a FedAvg round takes under a second with 20 clients. Geometric evidence is the cheapest variant, and functional evidence adds the proxy evaluations. With partial participation the server step handles only the sampled clients: with 100 CIFAR-10 clients at 10% participation, global FedCRF adds 29–39% to a 1.2 s round, about as much as FedLWS (+28%) and Zeno (+34%) and less than trimmed mean (+45%), RFA (+138%), and FedLAW (+149%).

Compared with the other server-side methods, global FedCRF costs about as much as FedLWS, FLTrust, and Zeno and less than the robust aggregators based on sorting or iterative estimation: coordinate-wise median and trimmed mean (+24–25% on CIFAR-10 with 20 clients, +84–85% with 100, and +176–344% on FEMNIST), Multi-Krum (+22%, +62%, and +160–193%), and the geometric median and RFA (+38%, +103–104%, and +394–549%). FedLAW, which optimizes its weights on the proxy set, adds 42–333%. FedCDA's search takes about 14 minutes per round with 20 clients on CIFAR-10 and FEMNIST and more than four hours with 100, a factor of 115–4407 over FedAvg.

Clustered FedCRF, which trains one model per group after the warm-up, adds 10% on CIFAR-10 with 20 clients, 28% with 100, and 84–90% on FEMNIST, less than IFCA (+44–48% and +101–102%), FedSoft (+103–107% and +142–144%), and FedRC (+146% and +178–184%), which evaluate or train several models on every client, and comparable to CFL and FLUX. Personalized FedCRF adds 27% and 91% on CIFAR-10 with 20 and 100 clients, less than Ditto (+159% and +172%), FedFomo (+112% and +316%), and, with 20 clients, pFedGraph (+42%; +98% with 100), and 254–288% on FEMNIST, more than Ditto and pFedGraph (+184–230%) but far less than FedFomo ($\times$12–14). With the heavier EfficientNet-B0 on Fed-ISIC2019, all FedCRF variants add at most 6% per round.

H RELIABILITY DIAGNOSTICS

Figure 2 tracks mean inferred reliability separately for attackers or corrupted clients and the remaining clients on CIFAR-10.

Table 12: Computational cost relative to FedAvg (↓): median wall-clock time of a training round (local training, server step, and evaluation) on one RTX 4090, rounds 2–4 of one seed without attack. The first row is FedAvg's time per round; the others give the increase over it, and values above +1000% are given as a factor. All clients participate in every round; with 100 clients, FedCDA searches its cache in 15 batches instead of 3, so that each batch keeps about seven clients.

Method	CIFAR-10, K=20	CIFAR-10, K=100	FEMNIST, K=20	FEMNIST, K=100	Fed-ISIC2019, K=18
FedAvg (time per round)	7.35 s	10.20 s	0.75 s	3.56 s	38.95 s
FedCRF					
FedCRF_G	+10%	+29%	+104%	+95%	+1%
FedCRF_F	+12%	+37%	+121%	+144%	+3%
FedCRF_{G+F}	+13%	+40%	+126%	+148%	+4%
FedCRF_G, clustered	+10%	+28%	+90%	+84%	+1%
FedCRF_{G+F}, personalized (s=10)	+27%	+91%	+254%	+288%	+6%
Global aggregation					
FedProx	+49%	+38%	+19%	+16%	+15%
FedNova	+5%	+11%	+53%	+48%	+0%
FedLWS	+10%	+28%	+94%	+73%	+1%
FedLAW	+42%	+129%	+333%	+242%	+70%
FedCDA	×115	×1535	×1127	×4407	+594%
Robust aggregation					
Coord. median	+25%	+85%	+179%	+344%	+2%
Trimmed mean	+24%	+84%	+176%	+342%	+2%
Krum	+20%	+59%	+177%	+146%	+2%
Multi-Krum	+22%	+62%	+193%	+160%	+2%
RFA	+38%	+103%	+549%	+542%	+10%
Geom. median	+38%	+104%	+492%	+394%	+9%
Fed-NGA	+23%	+73%	+217%	+188%	+2%
FLTrust	+18%	+36%	+91%	+81%	+7%
Zeno	+11%	+35%	+105%	+95%	+3%
Clustered FL					
IFCA	+44%	+48%	+101%	+102%	+146%
CFL	+15%	+40%	+137%	+126%	+2%
FedSoft	+107%	+103%	+144%	+142%	+162%
FedRC	+146%	+146%	+178%	+184%	+588%
FLUX	+9%	+17%	+78%	+46%	+7%
Personalized FL					
Local only	+5%	+13%	+19%	+18%	+1%
FedAvg + FT	+4%	+9%	+30%	+31%	+0%
Ditto	+159%	+172%	+184%	+187%	+116%
pFedGraph	+42%	+98%	+230%	+188%	+6%
FedFomo	+112%	+316%	×12	×14	+29%

Geometric evidence assigns near-zero reliability to sign-flipping and model-replacement attackers from the first round. ALIE is designed to evade such statistics, since its updates stay within the spread of the benign ones; there, the pairwise potentials strengthen geometric evidence (Table 13), and functional evidence suppresses all three attacks: ALIE attackers separate after two rounds and reach near-zero reliability by round seven, while benign clients settle at 0.4–0.5.

For corrupted data, geometric evidence drives the reliability of label-noisy clients to near zero, by round eight at 80% label noise and by round fifteen at 50%, and lowers input-noisy clients to 0.15–0.2 against about 0.3 for clean clients. Functional evidence keeps label-noisy clients at 0.15–0.25, since their updates still improve the proxy loss early in training; the two evidence sources are therefore complementary.

Mixed evidence combines both sources and lowers a client's reliability to around 0.15–0.2 when either source flags it. Inferred reliability also tracks each client's actual contribution: it correlates with per-client proxy-loss improvement on CIFAR-10 and high-severity MNIST (Pearson $r = 0.44$–0.74), while on IID and medium-severity MNIST, where clients differ little in quality, there is little variation to track ($r \approx 0$).

Pairwise potentials under ALIE. ALIE is the attack on which unary geometric evidence alone is least informative, which makes it a direct test of the importance of pairwise potentials. Table 13 runs FedCRF_G on CIFAR-10 under ALIE with ten seeds and three coupling strengths: $\lambda = 0$ (unary potentials only), and $\lambda = \{0.5, 2\}$. The pairwise term consistently shifts weight from the attackers to the benign clients: the attackers' share of the aggregate falls from 0.246 without coupling to 0.231 at $\lambda = 0.5$ and to 0.184 at $\lambda = 2$, and it is lower in every one of the ten seeds (the per-seed ranges 0.242–0.250, 0.223–0.240, and 0.181–0.186 do not overlap). Lower attacker weight translates into

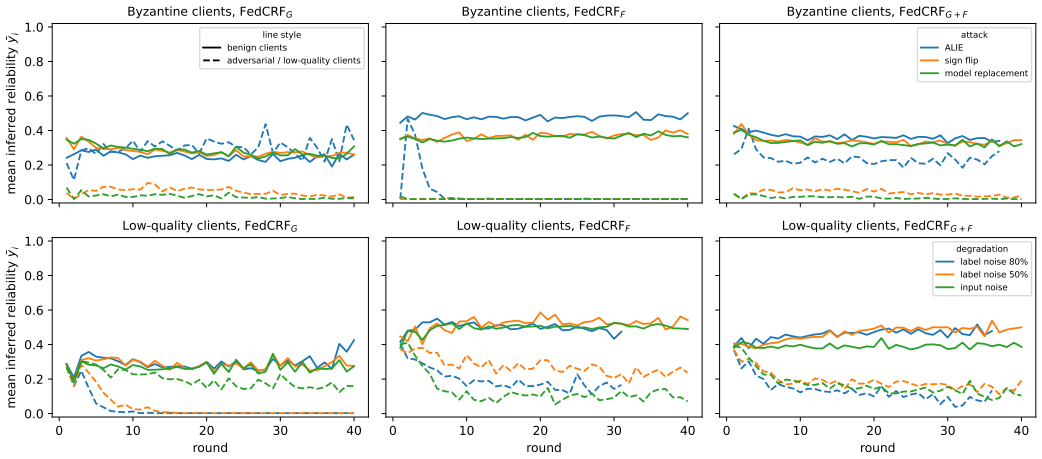


Figure 2: Mean inferred reliability on CIFAR-10 over 3 seeds. Dashed lines denote malicious clients (top, 20%) or clients with corrupted training data (bottom, 30%); solid lines denote the remaining clients. Columns compare geometric, functional, and mixed evidence.

more accurate and more stable training: with $\lambda = 2$, mean accuracy rises by 2.0 points (from 73.8% to 75.8%), the standard deviation across seeds falls from 4.1 to 1.8 points, and the worst seed improves from 63.4% to 73.0%. The default $\lambda = 0.5$ already raises the worst seed to 69.3% and lowers the standard deviation to 3.0 points. Coupling clients through their affinities thus lets the benign majority support one another's reliability, which a per-client score alone cannot do.

Table 13: Pairwise coupling under ALIE: FedCRF$_G$ on CIFAR-10 ($P(y)$ medium, 20% malicious clients), 10 seeds. Clean accuracy (%, mean $\pm$ standard deviation and worst seed) and the attackers' share of the aggregation weight, averaged over rounds.

Coupling	accuracy ↑	worst seed ↑	attacker weight ↓
$\lambda = 0$ (unary potentials only)	73.8$\pm$4.1	63.4	0.246
$\lambda = 0.5$	73.7$\pm$3.0	69.3	0.231
$\lambda = 2$	75.8$\pm$1.8	73.0	0.184