SketchGuard: Scaling Byzantine-Robust Decentralized Federated Learning via Sketch-Based Screening

Murtaza Rangwala, Farag Azzedin, Richard O. Sinnott, and Rajkumar Buyya

***Abstract*—Decentralized Federated Learning (DFL) enables privacy-preserving collaborative training without centralized servers but remains vulnerable to Byzantine attacks. Existing Byzantine-robust defenses are predicated on exchanging full, high-dimensional model vectors with every neighbor before filtering, an $O(d|\mathcal{N}_i|)$ communication cost incurred regardless of how many neighbors are ultimately rejected. This design choice is sustainable in small-scale experimental settings but becomes a fundamental barrier to deployment as network scale or model size grows. We propose SKETCHGUARD, a framework that decouples Byzantine filtering from aggregation via sketch-based screening. Each client compresses its d-dimensional model to a k-dimensional Count Sketch ($k \ll d$), exchanges only sketches for neighbor screening, and fetches full models exclusively from accepted neighbors. This eliminates the pre-filtering communication waste of existing defenses: rejected Byzantine neighbors incur only $O(k)$ sketch cost rather than $O(d)$ full-model cost. Communication savings therefore scale with the Byzantine rejection rate: negligible extra overhead in benign conditions, rising to 50–70% total savings when 50–70% of neighbors are rejected. We prove convergence in both strongly convex and non-convex settings, establishing that Count Sketch's distance-preservation guarantee causes sketch-based filtering to deviate from full-precision filtering by at most a $(1+O(\epsilon))$ factor in the effective threshold, a gap that can be made arbitrarily small. Experiments across three non-IID federated benchmarks, five network topologies, and four attack types confirm that SKETCHGUARD matches state-of-the-art robustness (mean TER deviation $\leq$0.5 percentage points) while reducing computation by up to 82%, with robustness remaining stable across compression ratios up to 13,000:1.**

***Index Terms*—Decentralized federated learning, Byzantine robustness, Count Sketch, communication efficiency, adversarial machine learning, distributed optimization.**

I. INTRODUCTION

FEDERATED Learning (FL) enables collaborative training of AI models over distributed data while preserving privacy by keeping raw data local [1]. The canonical server-assisted architecture, however, centralizes aggregation of model parameters, creating a single point of failure, a communication bottleneck, and a trust assumption that a central coordinator will behave honestly [2]. These limitations have motivated Decentralized Federated Learning (DFL), where clients exchange model updates directly in a peer-to-peer manner over a communication graph, eliminating any central authority and improving fault resilience [3], [4].

M. Rangwala, R. O. Sinnott, and R. Buyya are with the School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC, Australia.

F. Azzedin is with the Department of Information and Computer Science, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

Despite its architectural advantages, DFL introduces a distinctive adversarial challenge: *Byzantine robustness*. Without a central server to act as a gatekeeper, each client must independently decide which of its neighbors' updates to trust and incorporate. A malicious neighbor can submit arbitrary or carefully crafted updates to poison training, induce consensus drift, or embed hidden functionality [5], [6]. The problem is compounded by the graph-limited view of each client: unlike centralized FL where a global robust aggregation rule can reason over all participants simultaneously, each DFL client sees only its local neighborhood, often over non-IID data and time-varying connections.

The dominant defense paradigm in DFL is *local-consistency filtering*: a client accepts neighbor j's update only if it is sufficiently similar to the client's own current model [2], [7]–[10]. This approach has sound theoretical grounding—methods such as BALANCE [2] and SCCLIP [8] provide convergence guarantees in both strongly convex and non-convex settings—but it rests on a costly architectural assumption: *every neighbor's full model must be received and compared before any filtering decision can be made*. A client with $|\mathcal{N}_i|$ neighbors must therefore receive $O(d \cdot |\mathcal{N}_i|)$ parameters per round regardless of how many will ultimately be rejected. Crucially, this pre-filtering cost is paid in full even when all rejected neighbors could have been identified from a cheap compressed representation alone.

This is not merely a performance overhead. As network size grows—either because the DFL graph is larger or because honest clients are more densely connected—the pre-filtering communication cost scales linearly with degree, independent of attack intensity. For large models (which are increasingly common in FL deployments [11]), this scaling makes existing Byzantine-robust DFL schemes impractical: a node with 100 neighbors and a 60M-parameter model must exchange 24GB of data per round before a single filtering decision is made. Even in moderate-scale deployments over bandwidth-limited links, such as IoT networks, mobile edge computing clusters, or multi-institutional federated systems with metered inter-site connections [3], [12], this bottleneck is a fundamental barrier rather than an engineering inconvenience.

Sketch-based data structures offer a principled path forward. Count Sketch [13] can compress a d-dimensional vector to a k-dimensional summary ($k \ll d$) using random hash and sign projections, with the critical property that it *approximately preserves Euclidean distances*. If local-consistency filtering is based on distance comparisons, and if those distances

are approximately preserved under sketching, then filtering decisions made on compressed sketches should closely mirror decisions made on full models—enabling us to defer the expensive full-model exchange until *after* the filtering step, and only for accepted neighbors.

In this paper, we formalize and analyze this idea in SKETCHGUARD, a framework for Byzantine-robust DFL that decouples filtering from aggregation through sketch-based neighbor screening. SKETCHGUARD is a *general wrapper*: it is applicable to any similarity-based Byzantine defense that operates on Euclidean distances [2], [8], [14], [15]. We instantiate it with the state-of-the-art BALANCE aggregation rule [2] for theoretical analysis and empirical evaluation. Our main contributions are:

- **Algorithm.** We propose SKETCHGUARD, which reduces per-round communication from $O(d|\mathcal{N}_i|)$ to $O(k|\mathcal{N}_i| + d|\mathcal{S}_i|)$ by exchanging sketches for filtering and full models only for accepted neighbors, with a sketch-recomputation step that prevents adversaries from exploiting the two-phase exchange.
- **Theory.** We prove that Count Sketch's distance-preservation guarantee implies that sketch-based filtering deviates from full-precision filtering by at most a $\sqrt{(1+\epsilon)/(1-\epsilon)}$ factor in the effective threshold, and derive convergence rates in both strongly convex and non-convex settings that match optimal rates up to this controlled factor.
- **Experiments.** Across three non-IID federated benchmarks (FEMNIST, CelebA, Sent140), five network topologies, and four attack types (directed deviation, Gaussian, Krum, backdoor), SKETCHGUARD matches state-of-the-art robustness within 0.5 percentage points of Test Error Rate (TER) while reducing computation by up to 82% and communication by 50–70% under adversarial conditions. In benign conditions, sketch overhead is negligible (<0.02% of a full-model exchange) and the full-model fetch cost is unchanged since most neighbors are accepted. Robustness is stable across sketch sizes spanning compression ratios from 7:1 to 13,000:1.

II. PRELIMINARIES AND RELATED WORK

A. DFL Problem Formulation and Protocol

Consider n clients connected by an undirected graph $G = (V, E)$, where each client $i \in V$ possesses a private dataset $\mathcal{D}_i$ and maintains a local model $\mathbf{w}_i \in \mathbb{R}^d$. The collective objective is to minimize the average empirical loss:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}), \tag{1}$$

where $f_i(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_i}[\ell(\mathbf{w};\mathbf{x},y)]$ is the expected loss over client i's data distribution. The DFL protocol alternates between two phases. First, clients perform local gradient updates:

$$\mathbf{w}_i^{t+1/2} = \mathbf{w}_i^t - \eta\nabla f_i(\mathbf{w}_i^t). \tag{2}$$

Then, each client aggregates neighbor models according to:

$$\mathbf{w}_i^{t+1} = \alpha\mathbf{w}_i^{t+1/2} + (1-\alpha)\cdot \mathrm{AGG}_i\Big(\{\mathbf{w}_j^{t+1/2} : j \in \mathcal{N}_i\}\Big), \tag{3}$$

where $\mathcal{N}_i$ denotes the neighbors of client i and AGG_i is a local aggregation function. In the Byzantine-robust setting, AGG_i filters incoming models before combining them.

B. Byzantine Attack Models

We consider f-Byzantine adversaries that control up to f clients in each neighborhood. Byzantine clients can deviate arbitrarily from the protocol; they are aware of the aggregation mechanism and may collude. The convergence theorems in Section IV require $f < |\mathcal{N}_i|/2$, which is the standard assumption under which honest-majority filtering is theoretically tractable [2], [8]. Our experiments deliberately exceed this bound, evaluating Byzantine fractions up to 80%, to empirically characterize behavior under adversarial majorities, where theoretical guarantees no longer apply but practical robustness may still hold. We evaluate against four attack strategies that span the threat spectrum from untargeted disruption to adaptive manipulation and covert backdoor injection.

Directed Deviation (DD). An optimization-based attack [16] that crafts malicious updates to maximally displace the aggregated model away from the honest gradient direction:

$$\hat{\mathbf{w}}_j = \bar{\mathbf{w}}_{\mathcal{H}} - \lambda \cdot \frac{\bar{\mathbf{w}}_{\mathcal{H}} - \mathbf{w}_i^t}{\|\bar{\mathbf{w}}_{\mathcal{H}} - \mathbf{w}_i^t\|}, \tag{4}$$

where $\bar{\mathbf{w}}_{\mathcal{H}}$ is the mean of honest neighbors' updates (estimated by the attacker from prior rounds) and λ controls attack magnitude. We follow [2] in setting λ proportional to the standard deviation of honest updates, which produces adversarial models that are plausibly close in distance to their honest counterparts and therefore harder for distance-based filters to reject.

Gaussian Attack. Injects i.i.d. noise $\hat{\mathbf{w}}_j \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with $\sigma^2 = 200$ [5]. This models an adversary with no knowledge of the aggregation mechanism or honest updates, whose goal is blunt disruption of convergence. High variance ensures the injected models are far from honest neighbors in Euclidean distance, making this the easiest attack for similarity-based filters to detect.

Krum Attack. A targeted evasion of the Krum defense [16]. The attacker first selects a primary adversarial model $\hat{\mathbf{w}}^*$ pointing away from the honest gradient, then generates $f-1$ supporting Byzantine models clustered tightly around $\hat{\mathbf{w}}^*$:

$$\hat{\mathbf{w}}_j = \hat{\mathbf{w}}^* + \boldsymbol{\xi}_j, \quad \boldsymbol{\xi}_j \sim \mathcal{N}(\mathbf{0}, \sigma_s^2\mathbf{I}), \tag{5}$$

with σ_s^2 chosen small enough that the cluster is mutually close. The intent is for Krum's scoring function, which selects the model with the smallest sum of distances to its nearest neighbors, to score $\hat{\mathbf{w}}^*$ highly due to its supporting cluster. We include this attack both to stress-test Krum as a baseline and to evaluate whether sketch-based filtering inherits any vulnerability to distance-structure manipulation.

Backdoor Attack. Implants hidden behavior in clients' models without degrading main-task accuracy [17]. Each Byzantine client injects a trigger pattern into a subset of its training data and trains toward a target label on triggered inputs. To persist through repeated local aggregations across the graph, malicious clients apply a scaling factor $\gamma_{\mathrm{bd}} > 1$ to their model updates

before submission, amplifying the backdoor signal relative to honest contributions. Specifically:

$$\hat{\mathbf{w}}_j = \gamma_{\mathrm{bd}} \cdot \mathbf{w}_j^{\mathrm{bd}}, \tag{6}$$

where $\mathbf{w}_j^{\mathrm{bd}}$ is the locally trained model on the poisoned dataset and γ_{bd} is set to overcome honest averaging, following the model-replacement methodology of [17]. Trigger specifications are given in the supplemental material.

More sophisticated adversaries operate adaptively, tuning their attack based on observed defenses [18], and collusive strategies can coordinate deviations across multiple Byzantine nodes to overwhelm local neighborhood views [19]. Our focus on these four canonical attacks reflects standard practice in the DFL robustness literature [2], [7], [8]; adaptive attack evaluation against sketch-specific strategies is discussed in Section VI.

C. Byzantine-Robust DFL Defenses

Existing DFL defenses have converged on similarity-based neighbor filtering as the dominant paradigm, with methods differing primarily in their filtering criteria and aggregation mechanisms.

UBAR [7] uses two-stage neighbor selection based on Euclidean distance thresholding followed by loss-value filtering, with surviving neighbors averaged. It requires neither knowledge of the Byzantine fraction nor global connectivity but lacks formal convergence guarantees.

LEARN [10] employs multiple rounds of model exchange per training iteration with trimmed-mean aggregation. The multi-round communication structure makes it particularly expensive and exacerbates the scalability problem that motivates our work.

SCCLIP [8] applies self-centered clipping: each received model is clipped toward the client's own model by a fixed radius. This provides non-convex convergence guarantees without requiring knowledge of the Byzantine fraction, but the clipping radius must be set conservatively, and the method still requires receiving all neighbors' full models before clipping.

BALANCE [2] applies an adaptive exponentially decaying similarity threshold—the same structure we adopt in SKETCHGUARD's filtering step—and provides convergence guarantees in both strongly convex and non-convex settings without assumptions on Byzantine fraction knowledge or graph completeness. It represents the current state of the art in Byzantine-robust DFL aggregation.

WFAgg [9] proposes a multi-filter approach designed for dynamic and time-varying topologies, maintaining similar theoretical limitations to BALANCE in terms of communication cost.

A key point emphasized in Table I is that *all* of these methods, including the theoretically strongest, require receiving every neighbor's full d-dimensional model before making any filtering decision. SKETCHGUARD is the only method in this landscape that provides both convex and non-convex convergence guarantees *and* reduces this pre-filtering cost.

TABLE I
COMPARISON OF BYZANTINE-ROBUST DECENTRALIZED AGGREGATION ALGORITHMS. "CONV." / "NON-CV." INDICATE CONVERGENCE GUARANTEES; "NO c_i" MEANS NO NEED TO KNOW THE COMPROMISED NODE RATIO; "NO CMPL." MEANS NO COMPLETE-GRAPH ASSUMPTION; "SCALABILITY" REFLECTS COMMUNICATION EFFICIENCY AT SCALE (HIGHER IS BETTER).

Algorithm	Conv.	Non-cv.	No c_i	No Cmpl.	Scalability
UBAR [7]	–	–	–	–	Med.
LEARN [10]	–	✓	–	–	Low
SCCLIP [8]	–	✓	✓	✓	Med.
BALANCE [2]	✓	✓	✓	✓	Med.
WFAgg [9]	–	–	✓	✓	Med.
SKETCHGUARD	✓	✓	✓	✓	**High**

D. Compression Techniques in Federated Learning

Communication efficiency in FL has been approached through quantization [20], sparsification [21], [22], and low-rank approximation [23], [24]. These methods reduce the cost of transmitting model updates but were designed for the centralized FL setting and do not directly address Byzantine robustness.

Several works combine compression with Byzantine robustness in *centralized* FL [25], [26]. However, centralized robust compression methods rely on the parameter server's global view of all client updates; the server can apply robust aggregation rules such as geometric median or trimmed mean over the full client population before or after decompression. In fully decentralized FL, no such global view exists: each client aggregates from a local neighborhood only, and the filtering must be performed locally using only the information available at that node. This structural difference means that centralized robust compression schemes do not transfer to the decentralized setting, and dedicated methods are required.

FetchSGD [27] is the closest prior work to ours: it applies Count Sketch to compress gradients in the centralized FL setting and uses the sketch's distance-preservation properties for efficient aggregation. We extend this intuition to the Byzantine-robust *decentralized* setting, where sketches must serve as proxies for filtering decisions rather than for gradient reconstruction.

E. Count Sketch

Count Sketch [13] provides a randomized linear projection with properties that make it particularly suitable for Byzantine-robust neighbor screening. Given $\mathbf{w} \in \mathbb{R}^d$, a Count Sketch of size $k \ll d$ is constructed using a hash function $h : [d] \to [k]$ and a sign function $s : [d] \to \{-1, +1\}$, both drawn uniformly at random:

$$\mathrm{CS}(\mathbf{w})[b] = \sum_{i:\, h(i)=b} s(i)\, w_i, \quad b = 1, \ldots, k. \tag{7}$$

Three properties make Count Sketch well-suited to our problem. First, *linearity*: $\mathrm{CS}(\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha\,\mathrm{CS}(\mathbf{u}) + \beta\,\mathrm{CS}(\mathbf{v})$,

which ensures that all clients using the same hash functions compress consistently—a prerequisite for distance comparisons across clients to be meaningful. Second, *unbiasedness*: $\mathbb{E}[\mathrm{CS}(\mathbf{w})[b] \cdot s(i)] = w_i$ for any i with $h(i) = b$, providing formal coordinate recovery guarantees [13], [27]. Third, and most critical for our application, Count Sketch *approximately preserves Euclidean distances*:

Lemma II.1 (Distance Preservation [13])**.** *For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ and sketch size $k = O(\epsilon^{-2}\log(1/\zeta))$, with probability at least $1-\zeta$:*

$$(1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \le \|\mathrm{CS}(\mathbf{u})-\mathrm{CS}(\mathbf{v})\|^2 \le (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2. \quad (8)$$

This lemma is the key technical bridge between sketch compression and Byzantine filtering: if a neighbor's model is far from the client's own model in full-precision space, it will also be far in sketch space with high probability, and the filtering decision will agree. Section IV formalizes exactly how much the filtering threshold must be adjusted to account for the approximation gap.

III. SketchGuard: Scalable Robust Aggregation

A. Design Rationale

The central observation motivating SketchGuard is that in existing local-consistency defenses, the communication cost and the filtering decision are coupled by design: a client must receive a neighbor's full model before it can compute the distance used to decide whether to accept that neighbor. This coupling is not logically necessary. Distance comparisons only require that distances be *approximately preserved*—which sketch compression guarantees—while full model fidelity is needed only for the aggregation step that follows filtering.

SketchGuard breaks this coupling by splitting each round into two communication phases: a cheap sketch-exchange phase used for filtering, and a selective full-model fetch phase restricted to accepted neighbors. The result is that the expensive $O(d)$-per-neighbor communication is incurred only for neighbors that pass the filter, while the cost for rejected Byzantine neighbors is reduced to the much cheaper $O(k)$ sketch exchange.

B. Protocol Description

At each DFL round t, client i executes four phases, detailed in Algorithm 1 and illustrated in Fig. 1.

Phase 1: Local Training. Client i applies one step of stochastic gradient descent to its local dataset as in (2), producing the half-updated model $\mathbf{w}_i^{t+1/2}$.

Phase 2: Sketch Exchange. Each client computes and broadcasts its sketch: $\mathbf{s}_i^{t+1/2} = \mathrm{CS}(\mathbf{w}_i^{t+1/2}) \in \mathbb{R}^k$. All clients in the neighborhood $\mathcal{N}_i$ receive each other's sketches. Because $k \ll d$, this phase transmits $O(k \cdot |\mathcal{N}_i|)$ parameters in total—orders of magnitude less than the $O(d \cdot |\mathcal{N}_i|)$ cost of full-model exchange.

Algorithm 1 SketchGuard: Robust Aggregation via Adaptive Sketch-Based Filtering

Require: Local data $\mathcal{D}_i$, neighbors $\mathcal{N}_i$, parameters γ, κ, α, sketch size k
Ensure: Updated model $\mathbf{w}_i^{t+1}$
1: $\mathbf{w}_i^{t+1/2} \leftarrow \mathbf{w}_i^t - \eta \mathbf{g}(\mathbf{w}_i^t)$
2: $\mathbf{s}_i^{t+1/2} \leftarrow \mathrm{CS}(\mathbf{w}_i^{t+1/2})$
3: Broadcast $\mathbf{s}_i^{t+1/2}$ to all $j \in \mathcal{N}_i$; receive $\mathbf{s}_j^{t+1/2}$ from all $j \in \mathcal{N}_i$
4: $\tau \leftarrow \gamma \exp(-\kappa t/T)\|\mathbf{s}_i^{t+1/2}\|$
5: $\mathcal{S}_i^t \leftarrow \{j \in \mathcal{N}_i : \|\mathbf{s}_i^{t+1/2} - \mathbf{s}_j^{t+1/2}\| \le \tau\}$
6: **if** $|\mathcal{S}_i^t| = 0$ and $|\mathcal{N}_i| > 0$ **then**
7: $\quad \mathcal{S}_i^t \leftarrow \{\arg\min_{j \in \mathcal{N}_i} \|\mathbf{s}_i^{t+1/2} - \mathbf{s}_j^{t+1/2}\|\}$
8: **end if**
9: Fetch full models $\{\mathbf{w}_j^{t+1/2}\}_{j \in \mathcal{S}_i^t}$ from accepted neighbors
10: For each $j \in \mathcal{S}_i^t$: verify $\mathrm{CS}(\mathbf{w}_j^{t+1/2}) = \mathbf{s}_j^{t+1/2}$; discard if mismatch
11: $\mathbf{w}_i^{t+1} \leftarrow \alpha \mathbf{w}_i^{t+1/2} + \frac{1-\alpha}{|\mathcal{S}_i^t|} \sum_{j \in \mathcal{S}_i^t} \mathbf{w}_j^{t+1/2}$
12: **return** $\mathbf{w}_i^{t+1}$

Phase 3: Adaptive Sketch-Domain Filtering. Client i accepts neighbor j if the sketch-domain distance satisfies an adaptively decaying threshold:

$$\|\mathbf{s}_i^{t+1/2} - \mathbf{s}_j^{t+1/2}\| \le \gamma \exp(-\kappa t/T)\|\mathbf{s}_i^{t+1/2}\|, \quad (9)$$

where $\gamma > 0$ controls the base acceptance radius, $\kappa > 0$ controls exponential tightening over T total rounds, and the decay reflects the convergence of honest clients toward the optimum: as honest clients' models align, the acceptable spread among them narrows. The accepted set is $\mathcal{S}_i^t = \{j \in \mathcal{N}_i : (9) \text{ holds}\}$.

Remark III.1. When Count Sketch is used with approximation parameter ϵ, Lemma II.1 implies that a full-precision distance satisfying the threshold γ will have its sketch-domain counterpart satisfy at most $\gamma_{\text{eff}} = \gamma\sqrt{(1+\epsilon)/(1-\epsilon)}$. The convergence analysis in Section IV shows that this effective threshold replaces γ throughout the proof with no other structural changes.

Phase 4: Verified Model Aggregation. Full models $\{\mathbf{w}_j^{t+1/2}\}_{j \in \mathcal{S}_i^t}$ are fetched only from accepted neighbors. To guard against adversaries that pass the sketch filter but send a different full model, each received model is verified by recomputing its sketch and comparing against the sketch exchanged in Phase 2; any mismatch causes that neighbor to be discarded. Verified models are then aggregated:

$$\mathbf{w}_i^{t+1} = \alpha \mathbf{w}_i^{t+1/2} + \frac{(1-\alpha)}{|\mathcal{S}_i^t|} \sum_{j \in \mathcal{S}_i^t} \mathbf{w}_j^{t+1/2}, \quad (10)$$

where $\alpha \in [0, 1]$ balances self-reliance and collaboration.

Remark III.2. Lines 6–8 of Algorithm 1 handle the degenerate case $|\mathcal{S}_i^t| = 0$, which cannot occur under the $f < |\mathcal{N}_i|/2$ regime assumed by the theorems (at least one honest neighbor always satisfies the threshold). This fallback ensures the algorithm remains well-defined outside the theorems' regime, e.g., during early rounds with a poorly calibrated γ.

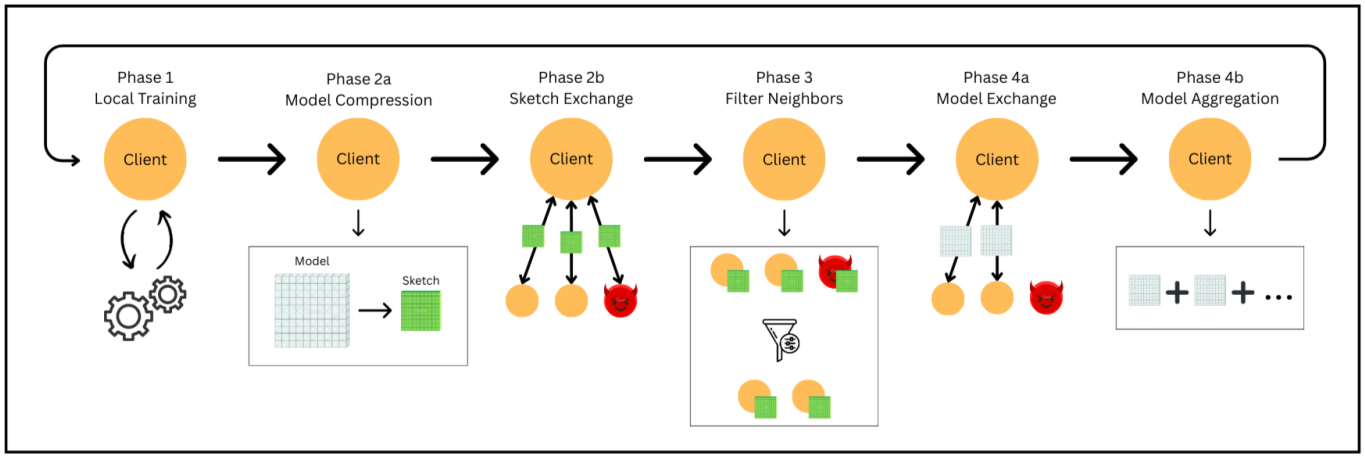

Fig. 1. The SKETCHGUARD protocol. In Phase 2, all neighbors exchange k-dimensional sketches ($k \ll d$). In Phase 3, each client computes sketch-domain distances and identifies accepted neighbors $\mathcal{S}_i^t$. In Phase 4, full d-dimensional models are fetched only from $\mathcal{S}_i^t$, and each is verified by recomputing its sketch before aggregation. Byzantine neighbors that fail the sketch-domain filter incur only $O(k)$ communication cost rather than $O(d)$.

TABLE II
PER-NODE-PER-ROUND COMPLEXITY COMPARISON. SOTA REFERS TO FULL-PRECISION SIMILARITY-BASED METHODS (BALANCE, SCCLIP, UBAR).

Phase	**SOTA**	**SketchGuard**
Local training	$O(d)$	$O(d)$
Sketch generation	–	$O(d)$
Neighbor screening	$O(d \cdot \lvert\mathcal{N}_i\rvert)$	$O(k \cdot \lvert\mathcal{N}_i\rvert)$
Verif. & aggregation	$O(d \cdot \lvert\mathcal{S}_i^t\rvert)$	$O(d \cdot \lvert\mathcal{S}_i^t\rvert)$
Total (compute)	$O(d(1 + \lvert\mathcal{N}_i\rvert + \lvert\mathcal{S}_i^t\rvert))$	$O(d(1 + \lvert\mathcal{S}_i^t\rvert) + k\lvert\mathcal{N}_i\rvert)$
Total (comm.)	$O(d \cdot \lvert\mathcal{N}_i\rvert)$	$O(k\lvert\mathcal{N}_i\rvert + d\lvert\mathcal{S}_i^t\rvert)$

C. Security of the Two-Phase Exchange

A natural concern with the two-phase design is that it introduces a new attack surface: a Byzantine neighbor could send a benign sketch in Phase 2 to pass the filter, then send a malicious full model in Phase 4. The sketch recomputation step in Line 9 of Algorithm 1 closes this gap. Because Count Sketch is a deterministic function of the model (for fixed hash functions), a neighbor cannot pass the Phase 2 filter with one model while submitting a different model in Phase 4 without causing a detectable mismatch. The only residual risk is a Byzantine neighbor that constructs a malicious model $\hat{\mathbf{w}}$ that (i) lies within the filtering threshold in sketch space and (ii) differs from an honest model. This risk is exactly what the filtering threshold γ_{eff} bounds—it is identical in character to the residual risk in full-precision filtering with threshold γ, as discussed in Remark IV.11.

D. Complexity Analysis

Table II summarizes per-node-per-round costs. Let d be the model dimension, $|\mathcal{N}_i|$ the neighbor count, k the sketch size, and $|\mathcal{S}_i^t| \leq |\mathcal{N}_i|$ the accepted count.

The savings have a multiplicative structure. When the compression ratio is d/k and the rejection rate is $1 - |\mathcal{S}_i^t|/|\mathcal{N}_i|$, the communication reduction factor relative to SOTA is approximately:

$$\frac{k|\mathcal{N}_i| + d|\mathcal{S}_i^t|}{d|\mathcal{N}_i|} = \frac{k}{d} + \frac{|\mathcal{S}_i^t|}{|\mathcal{N}_i|}. \tag{11}$$

With $k/d \approx 1/6600$ (our experimental setting) and $|\mathcal{S}_i^t|/|\mathcal{N}_i| \approx 0.5$ (50% Byzantine rejection), this ratio is approximately 0.5, yielding the empirically observed 50% communication reduction. Higher rejection rates and larger models amplify these savings further.

IV. CONVERGENCE ANALYSIS

We establish that SKETCHGUARD maintains the convergence guarantees of full-precision Byzantine-robust aggregation despite performing filtering in the compressed sketch domain. The proof strategy is: (1) use Lemma II.1 to translate sketch-domain filtering decisions into bounds on full-precision model differences; (2) show those bounds enter the convergence analysis identically to the full-precision case, but with γ replaced by γ_{eff}; (3) derive convergence rates by applying strong convexity or non-convexity arguments. Full proofs are provided in the supplemental material; we present the key steps here.

A. Technical Assumptions

Assumption IV.1 (Strong Convexity)**.** The population risk $F(\mathbf{w})$ is μ-strongly convex: $\forall\, \mathbf{w}_1, \mathbf{w}_2 \in \Theta$,

$$F(\mathbf{w}_1) + \langle \nabla F(\mathbf{w}_1), \mathbf{w}_2 - \mathbf{w}_1 \rangle + \frac{\mu}{2}\|\mathbf{w}_2 - \mathbf{w}_1\|^2 \leq F(\mathbf{w}_2).$$

Assumption IV.2 (Smoothness)**.** $F(\mathbf{w})$ is L-smooth: $\forall\, \mathbf{w}_1, \mathbf{w}_2 \in \Theta$, $\|\nabla F(\mathbf{w}_1) - \nabla F(\mathbf{w}_2)\| \leq L\|\mathbf{w}_1 - \mathbf{w}_2\|$.

Assumption IV.3 (Bounded Stochastic Gradient Variance)**.** For any honest client $i \in \mathcal{H}$, the stochastic gradient is unbiased with bounded variance: $\mathbb{E}[\mathbf{g}(\mathbf{w}_i)] = \nabla F(\mathbf{w}_i)$ and $\mathbb{E}[\|\mathbf{g}(\mathbf{w}_i) - \nabla F(\mathbf{w}_i)\|^2] \leq \delta^2$.

Assumption IV.4 (Bounded Parameters)**.** For any honest client $i \in \mathcal{H}$: $\|\mathbf{w}_i\| \leq \psi$ and $\|\nabla F(\mathbf{w}_i)\| \leq \rho$.

Assumption IV.5 (Graph Connectivity)**.** The subgraph $G_{\mathcal{H}}$ induced by honest clients remains connected throughout training.

Assumption IV.6 (Shared Hash Functions)**.** All clients use identical hash function h and sign function s. This is satisfied by seeding both with the model dimension d, which is shared implicitly among all clients training the same architecture.

Assumptions IV.1–IV.4 are standard in the Byzantine FL literature [2], [8]. Assumption IV.5 rules out partitioning attacks that isolate honest clients and is standard in DFL analyses [7]. Assumption IV.6 is required for sketch distances to be comparable across clients and is trivially satisfied in practice.

Remark IV.7 (Graph topology and the proof structure). The convergence bound is stated per client: it tracks each benign client i's own model $\mathbf{w}_i^t$ against $\mathbf{w}^*$ directly, without requiring a graph mixing or spectral gap argument. This is because the proof does not track a network-average model — the neighbor aggregation term is bounded locally at each round by $\gamma_{\text{eff}}\psi$ (derived in the Key Lemma below), and does not accumulate across rounds through graph propagation. The communication graph enters only through Assumption IV.5 (connectivity of $G_{\mathcal{H}}$), which ensures no honest client is permanently surrounded by Byzantine neighbors. Since the sketch modification affects only the filtering criterion that constructs $\mathcal{S}_i^t$ — and not the aggregation weights, graph structure, or any other protocol component — this connectivity argument applies to SKETCHGUARD without modification.

B. Key Lemma: From Sketch Filtering to Model-Space Bounds

The following argument, derived from Lemma II.1, is the core technical step connecting sketch-based filtering to the convergence analysis.

Lemma IV.8 (Sketch Filtering Implies Model-Space Bound)**.** *For any neighbor j accepted by the sketch-domain filter, the full-precision distance between j's model and client i's model satisfies:*

$$\|\mathbf{w}_j^{t+1/2} - \mathbf{w}_i^{t+1/2}\| \leq \gamma_{\text{eff}} \, \exp(-\kappa t/T) \, \|\mathbf{w}_i^{t+1/2}\|, \quad (12)$$

where $\gamma_{\text{eff}} = \gamma\sqrt{(1+\epsilon)/(1-\epsilon)}$ is the effective threshold under Count Sketch approximation parameter ϵ. Averaging over all accepted neighbors $\mathcal{S}_i^t$ and applying the bounded-parameters Assumption IV.4 ($\|\mathbf{w}_i\| \leq \psi$):

$$\left\| \frac{1}{|\mathcal{S}_i^t|} \sum_{j \in \mathcal{S}_i^t} \left(\mathbf{w}_j^{t+1/2} - \mathbf{w}_i^{t+1/2}\right) \right\| \leq \gamma_{\text{eff}} \, \psi. \quad (13)$$

Both bounds hold simultaneously for all rounds and all clients with probability at least $1 - \zeta_{\text{sys}}$, where $\zeta_{\text{sys}} \in (0,1)$ is the overall sketch failure probability, for sketch size:

$$k = O\left(\epsilon^{-2} \log\left(\frac{T \cdot n \cdot \Delta}{\zeta_{\text{sys}}}\right)\right),$$

with $\Delta = \max_i |\mathcal{N}_i|$ denoting the maximum node degree. This follows from a union bound over T rounds, n nodes, and Δ neighbors per node.

Proof sketch. Fix any $j \in \mathcal{S}_i^t$. The sketch filter accepted j, so:

$$\|\text{CS}(\mathbf{w}_i^{t+1/2}) - \text{CS}(\mathbf{w}_j^{t+1/2})\| \leq \gamma \, e^{-\kappa t/T} \|\text{CS}(\mathbf{w}_i^{t+1/2})\|.$$

Applying Lemma II.1: the left side satisfies $\|\text{CS}(\mathbf{w}_i) - \text{CS}(\mathbf{w}_j)\| \geq (1-\epsilon)^{1/2}\|\mathbf{w}_i - \mathbf{w}_j\|$, and the right side satisfies $\|\text{CS}(\mathbf{w}_i)\| \leq (1+\epsilon)^{1/2}\|\mathbf{w}_i\|$. Substituting and rearranging:

$$\begin{aligned}\|\mathbf{w}_i^{t+1/2} - \mathbf{w}_j^{t+1/2}\| &\leq \gamma\sqrt{(1+\epsilon)/(1-\epsilon)}\, e^{-\kappa t/T} \|\mathbf{w}_i^{t+1/2}\| \\ &= \gamma_{\text{eff}} \, e^{-\kappa t/T} \|\mathbf{w}_i^{t+1/2}\|.\end{aligned}$$

Averaging over $j \in \mathcal{S}_i^t$ and applying $\|\mathbf{w}_i^{t+1/2}\| \leq \psi$ yields (13). □

Bound (13) is structurally identical to the key bound in the full-precision analysis of BALANCE [2]—it differs only in $\gamma \to \gamma_{\text{eff}}$. This means the entire subsequent convergence argument from [2] carries through with this substitution, yielding the theorems below.

C. Main Convergence Results

Theorem IV.9 (Strongly Convex Convergence)**.** *Under Assumptions IV.1–IV.6 with learning rate $\eta \leq \min\{1/(4L), 1/\mu\}$, sketch failure probability $\zeta_{\text{sys}} \in (0,1)$, and sketch size $k = O(\epsilon^{-2}\log(Tn\Delta/\zeta_{\text{sys}}))$, the following holds with probability at least $1 - \zeta_{\text{sys}}$. After T rounds:*

$$\mathbb{E}[F(\mathbf{w}_i^T) - F(\mathbf{w}^*)] \leq \underbrace{(1-\mu\eta)^T \left[F(\mathbf{w}_i^0) - F(\mathbf{w}^*)\right]}_{\textit{geometric decay}} + \underbrace{\frac{2L\eta\delta^2}{\mu}}_{\textit{gradient noise}} + \underbrace{\frac{2\gamma_{\text{eff}}\rho\psi(1-\alpha)}{\mu\eta}}_{\textit{sketch approx. + heterogeneity}}. \quad (14)$$

Proof. See the supplemental material (Section S1). □

Theorem IV.10 (Non-Convex Convergence)**.** *Under Assumptions IV.2–IV.6 with the same parameter choices as Theorem IV.9, with probability at least $1 - \zeta_{\text{sys}}$:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F(\mathbf{w}_i^t)\|^2] \leq \underbrace{\frac{2[F(\mathbf{w}_i^0) - F^*]}{\eta T}}_{\textit{initialization gap}} + \underbrace{4L\eta\delta^2}_{\textit{gradient noise}} + \underbrace{\frac{4\gamma_{\text{eff}}\rho\psi(1-\alpha)}{\eta}}_{\textit{sketch approx. + heterogeneity}}. \quad (15)$$

Proof. See the supplemental material (Section S2). □

Both theorems match the optimal convergence rates for their respective settings [28]. We make three observations about the role of the sketch approximation.

Interpretability of terms. The third term in each bound captures two effects jointly: the sketch approximation (through γ_{eff}) and honest neighbor heterogeneity (through $\rho\psi$). When

TABLE III
THEORETICAL SKETCH SIZE $k = \lceil 25 \log(Tn\Delta/0.05) \rceil$ (USING $\epsilon^{-2} = 25$ FOR $\epsilon = 0.2$) FOR REPRESENTATIVE DEPLOYMENT SCALES. ALL VALUES ARE FAR BELOW TYPICAL MODEL DIMENSIONS d.

n	Δ	T	Theoretical k	Example d
20	19	50	≈ 230	6.6M (FEMNIST)
100	20	100	≈ 420	6.6M
500	30	200	≈ 540	60M (XLarge)
1000	50	500	≈ 630	60M

$\epsilon \to 0$, $\gamma_{\text{eff}} \to \gamma$ and the bound reduces exactly to the full-precision BALANCE result [2]. The mixing parameter α appears as $(1-\alpha)$: increasing α toward 1 down-weights neighbor influence, shrinking this term at the cost of slower consensus.

Approximation gap is small. For $\epsilon = 0.2$ (our experimental setting), $\gamma_{\text{eff}} = \gamma\sqrt{1.2/0.8} \approx 1.22\gamma$, meaning the effective filtering threshold is approximately 22% looser than in full-precision filtering. In practice, this translates to accepting a slightly wider band of neighbors; our experiments confirm this has no measurable effect on robustness because honest neighbors remain well within this band.

Sketch size scaling. The required sketch size $k = O(\epsilon^{-2}\log(Tn\Delta/\zeta_{\text{sys}}))$ is independent of model dimension d — this is the source of SKETCHGUARD's sub-linear scaling with model size, since as d grows the sketch-phase communication cost remains fixed while savings on pre-filtering communication grow linearly with d. However, the sketch size does grow logarithmically with n, T, and Δ. This dependence should be stated precisely: it is *dimension-free* but not *deployment-free*. To quantify the practical impact, Table III evaluates the theoretical k for representative settings at $\epsilon = 0.2$ and $\zeta_{\text{sys}} = 0.05$. The logarithmic growth is mild: moving from our experimental setting ($n = 20$, $\Delta = 19$, $T = 50$) to a large-scale deployment ($n = 1000$, $\Delta = 50$, $T = 500$) increases the required k by less than $3\times$, from roughly 230 to roughly 630. Both values remain orders of magnitude below any realistic d, preserving the compression benefit even at scale.

D. Convergence-Theoretic Robustness Guarantee

Remark IV.11 (Scope of the robustness guarantee). The convergence bounds in Theorems IV.9 and IV.10 imply a precise, but bounded, robustness guarantee. Specifically, the proof requires only one property of the filtering step: that the average deviation of accepted neighbors from client i is bounded by $\gamma_{\text{eff}}\psi$ per round (Lemma IV.8). Any Byzantine strategy that keeps accepted neighbors within this bound — whether in the full-precision or sketch setting — will produce the same convergence outcome. The sketch introduces a $\sqrt{(1+\epsilon)/(1-\epsilon)}$ expansion of the effective threshold relative to full-precision filtering, meaning the set of Byzantine models that can pass the sketch filter is slightly larger than those that pass the full-precision filter. This is the only robustness cost of compression, and it is quantified.

This guarantee is convergence-theoretic, not a full adversarial security proof. In particular, it does not rule out attack strategies that exploit the two-phase protocol structure beyond what the sketch-verification step in Phase 4 addresses. The verification step (Algorithm 1, Line 9) provably closes the specific gap of sending a benign sketch in Phase 2 and a different full model in Phase 4, since Count Sketch is a deterministic function of the model for fixed hash functions. Any remaining attack surface is bounded by the same γ_{eff} threshold that governs full-precision filtering — a Byzantine neighbor can only influence the aggregation if its model lies within the acceptance region, regardless of whether filtering was performed in sketch or full-precision space.

V. PERFORMANCE EVALUATION

A. Experimental Setup

Datasets and Models
We evaluate on three benchmarks from the LEAF federated learning suite [29]. **FEMNIST** is a 62-class handwritten character recognition task over 3,550 users with a CNN of 6.6M parameters. **CelebA** is a binary smile-classification task over 9,343 users with a LeNet-style CNN of 2.2M parameters. **Sent140** is a Twitter sentiment analysis task over 660,120 users with a two-layer LSTM of 1.2M parameters. These three datasets cover both image and text modalities with naturally non-IID user distributions, spanning three orders of magnitude in model size. Detailed architecture specifications are in the supplemental material.

Network Topologies
We evaluate five topologies: Ring (degree 2), Erdős-Rényi (ER) with $p \in \{0.2, 0.45, 0.6\}$, and Fully Connected. ER topologies are *dynamic*: edges are resampled each round, modeling realistic peer-to-peer networks with intermittent connectivity. Robustness experiments use 20-node networks; scalability experiments use k-regular graphs with node counts from 20 to 300. Full topology parameters are in the supplemental material.

Evaluation Metrics
Test Error Rate (TER) = $1 -$ test accuracy, averaged across honest clients. Lower is better.

Attack Success Rate (ASR): For backdoor attacks, the fraction of triggered test inputs classified as the attacker's target label. The random-chance baseline is $1/C$: 1.6% for FEMNIST (62 classes), 50% for CelebA and Sent140 (binary).

Per-Round Computation Time: Wall-clock time for the neighbor screening and aggregation step per client per round, on identical CPU hardware, excluding local training (which is constant across all methods).

Communication Overhead: Total floating-point parameters transmitted per client per round.

Baselines and Configuration
Baselines are D-FedAvg [4] (no Byzantine defense), KRUM [5], UBAR [7], and BALANCE [2]. Sketch sizes are $k = 1000$

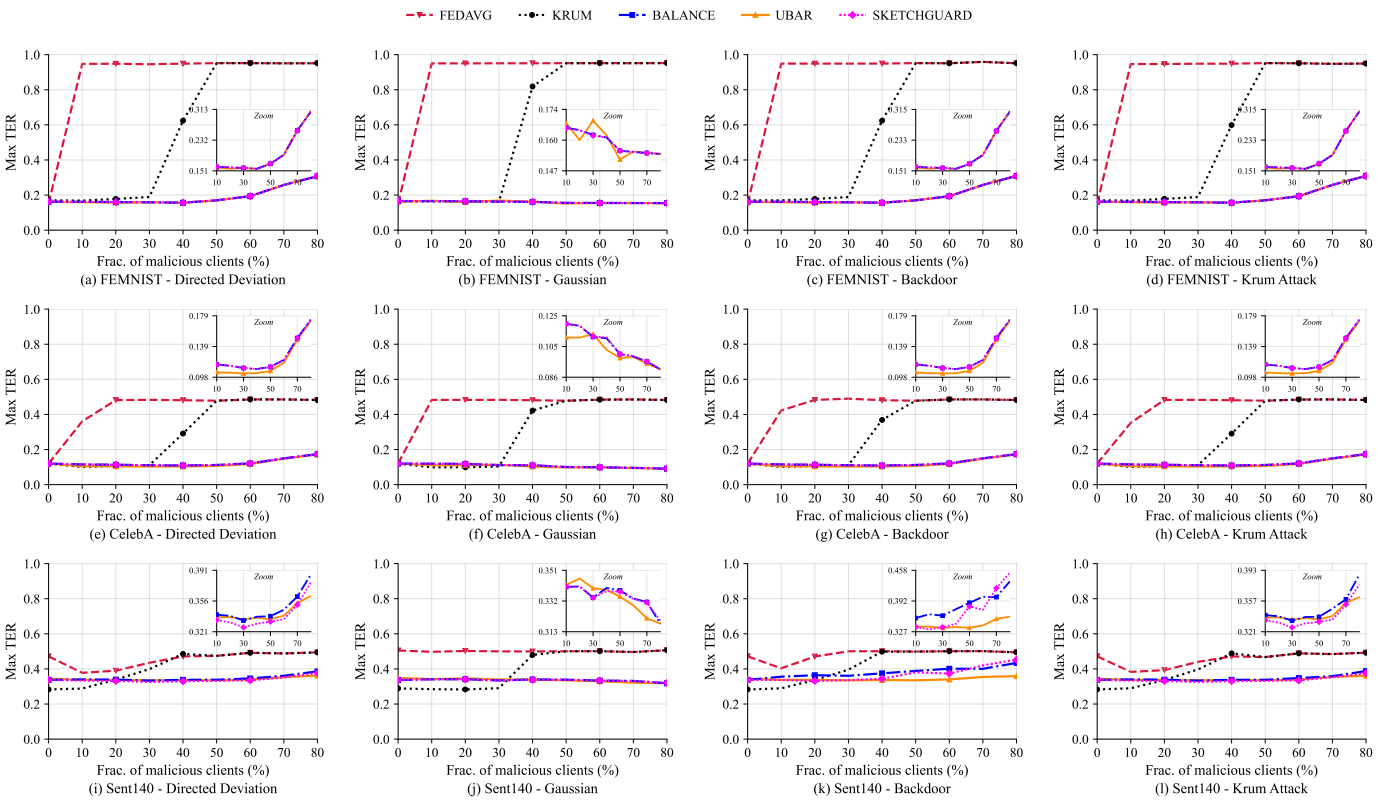

Fig. 2. Test Error Rate (TER) vs. fraction of Byzantine clients across datasets and attack types. SKETCHGUARD (SG) tracks BALANCE and UBAR throughout, confirming that sketch-domain filtering produces equivalent robustness to full-precision filtering across the full spectrum of attack intensities and types.

(FEMNIST, compression ratio ≈6600:1), $k = 350$ (CelebA, ≈6300:1), $k = 180$ (Sent140, ≈6700:1), giving $\epsilon \lesssim 0.2$ in all cases. SKETCHGUARD and BALANCE share identical threshold parameters $\gamma = 2.0$, $\kappa = 1.0$, $\alpha = 0.5$. All experiments run for 50 global rounds, 3 local epochs per round, averaged over 3 independent seeds. Complete hyperparameters are in the supplemental material.

B. Byzantine Robustness Evaluation

Fig. 2 shows TER as a function of Byzantine fraction across all datasets and attack types. The key finding is that SKETCHGUARD tracks BALANCE and UBAR throughout, with a mean absolute TER deviation of 0.38 percentage points versus BALANCE and 0.50 percentage points versus UBAR—well within the variance across seeds (typically ±0.3–0.5 percentage points). This confirms that the theoretical $O(\epsilon)$ threshold expansion has no practically meaningful effect on filtering decisions.

Against directed deviation and Krum attacks, all three similarity-based methods maintain TER below 20% on FEMNIST and below 13% on CelebA even at 80% Byzantine clients, while D-FedAvg and KRUM collapse to near-random accuracy (TER $>$ 60%). The Krum attack is particularly noteworthy: it is specifically designed to exploit distance-based filtering by constructing a cluster of mutually close Byzantine models. Despite this, all three similarity-based defenses remain robust, including SKETCHGUARD—the sketch compression does not amplify the vulnerability to this attack.

For Gaussian attacks, SKETCHGUARD matches BALANCE within 0.03 percentage points on average. Gaussian-injected models are far from honest clients in both full-precision and sketch space (high variance ensures this), making them the easiest to reject; all similarity-based methods handle this case well regardless of whether filtering is in full or sketch space.

For backdoor attacks, SKETCHGUARD achieves ASR within 3 percentage points of BALANCE and UBAR across all datasets. On FEMNIST, all three methods achieve ASR in the range 6.98–9.94% (versus the random baseline of 1.6%), reflecting that scale-amplified backdoor updates are detectable by similarity filters even in sketch space. On CelebA and Sent140, all robust methods substantially suppress ASR relative to D-FedAvg.

C. Computational Efficiency

Fig. 3 evaluates how per-round computation time scales with network size and model size on FEMNIST under 50% Byzantine clients.

Connectivity scaling (Fig. 3, top): As node degree increases from 16 to 299, BALANCE and UBAR scale linearly (dominated by their $O(d \cdot |\mathcal{N}_i|)$ screening cost), while SKETCHGUARD remains nearly flat at ≈0.35s, since its screening cost is $O(k \cdot |\mathcal{N}_i|)$ with $k \ll d$. At 299 neighbors, SKETCHGUARD is 60% faster than BALANCE (0.39s vs. 0.99s) and 82% faster than UBAR (0.39s vs. 2.14s).

Model size scaling (Fig. 3, bottom): From 220K to 60M parameters, SKETCHGUARD grows sub-linearly because the

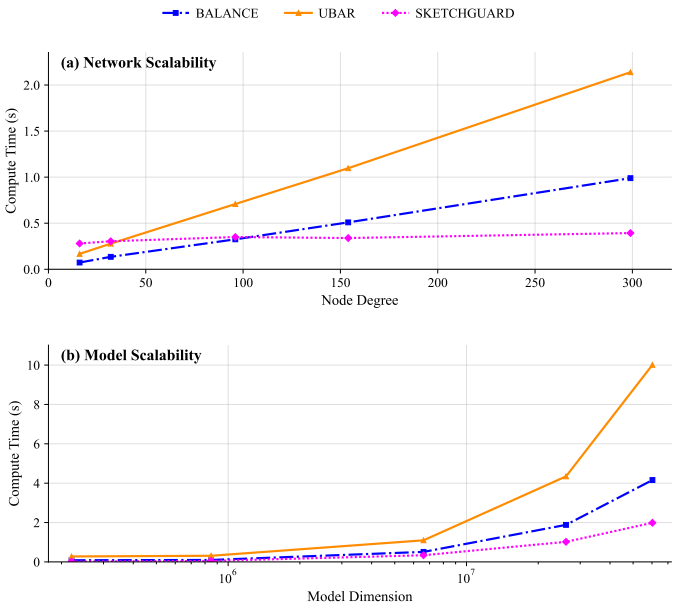

Fig. 3. Per-node computation time vs. network connectivity (top) and model size (bottom) on FEMNIST with 50% Byzantine clients under directed deviation. SKETCHGUARD's near-constant cost with connectivity and sub-linear growth with model size contrast with the linear scaling of BALANCE and UBAR.

TABLE IV
TER (%) ACROSS TOPOLOGIES, AVERAGED OVER DATASETS AND ATTACKS (EXCLUDING BACKDOOR).

Topology	Byz.%	UBAR	BALANCE	SG
Ring	20	19.0	19.9	19.9
	40	18.9	19.8	19.6
	60	30.2	30.6	30.3
	80	64.3	63.7	64.2
ER (p=0.2)	20	20.6	20.8	20.7
	40	20.2	20.4	20.3
	60	19.4	19.8	19.7
	80	18.8	20.4	20.0
ER (p=0.45)	20	20.1	20.4	20.3
	40	20.4	20.1	19.9
	60	19.6	19.8	19.5
	80	19.2	20.3	20.2
ER (p=0.6)	20	20.1	20.6	20.3
	40	20.1	20.2	20.1
	60	19.7	19.7	19.5
	80	19.1	19.2	19.0
Fully Conn.	20	20.0	20.5	20.3
	40	19.9	20.1	19.9
	60	19.5	19.7	19.5
	80	19.1	19.3	18.8

sketch size k is fixed by approximation parameters (ϵ, ζ_{sys}) rather than model dimension. At 60M parameters, SKETCHGUARD requires 2.0s vs. 4.2s (BALANCE, 52% reduction) and 10.0s (UBAR, 80% reduction).

D. Communication Efficiency

Per-round communication in SKETCHGUARD is $O(k|\mathcal{N}_i| + d|\mathcal{S}_i^t|)$, compared to $O(d|\mathcal{N}_i|)$ for full-precision baselines. The reduction depends on how many neighbors are rejected, so we report both operating conditions separately.

Benign conditions (no or few Byzantine clients, $|\mathcal{S}_i^t| \approx |\mathcal{N}_i|$): Almost all neighbors are accepted, so the full-model fetch cost is essentially unchanged relative to baselines. The sketch-exchange phase adds $O(k|\mathcal{N}_i|)$ overhead, which at our compression ratio of ≈6600:1 represents <0.02% of the full-model exchange cost and is negligible in practice. Total communication is therefore approximately the same as full-precision methods in benign conditions.

Adversarial conditions (50–70% Byzantine clients rejected): With $|\mathcal{S}_i^t| \approx 0.3$–$0.5 \cdot |\mathcal{N}_i|$, the communication reduction factor from Section III is $k/d + |\mathcal{S}_i^t|/|\mathcal{N}_i| \approx 0.30$–$0.50$, yielding 50–70% total communication savings relative to full-precision baselines. This is where SKETCHGUARD's decoupling of filtering from aggregation provides its primary communication benefit: rejected Byzantine neighbors incur only $O(k)$ sketch cost rather than $O(d)$ full-model cost.

E. Sensitivity Analysis

1) Sketch Size: Fig. 3 (bottom) already covers the effect of model size at a fixed sketch size. To evaluate the effect of varying k directly, we sweep k from 500 to 100,000 on FEMNIST and CelebA under 50% Byzantine clients with directed deviation. TER remains stable at 15.59% (FEMNIST) and 10.35% (CelebA) across all k values, including compression ratios exceeding 13,000:1. This insensitivity arises because, as Theorem IV.9 shows, the convergence bound depends on k only through $\gamma_{\text{eff}} = \gamma\sqrt{(1+\epsilon)/(1-\epsilon)}$, and this ratio changes slowly: even at 13,000:1 compression, ϵ remains small enough that γ_{eff} is within 5% of γ.

2) Network Topology: Table IV reports TER across all five topologies. SKETCHGUARD matches BALANCE within 0.5 percentage points in all cells. In well-connected topologies (ER $p \geq 0.45$, Fully Connected), all methods maintain TER below 20% even at 80% Byzantine clients, reflecting that denser honest subgraphs provide redundant paths for good information. In the Ring topology, all methods degrade at 60%+ Byzantine fractions due to the minimal honest neighborhood: with only 2 neighbors total, a single Byzantine node represents 50% of the neighborhood, making filtering inherently harder. This is a fundamental property of local aggregation under high attack rates in sparse graphs, not a weakness of any specific method.

The slight TER decrease at higher Byzantine fractions in well-connected topologies (visible at $p = 0.6$ and Fully Connected) is a known property of similarity-based filtering [2]: at high Byzantine fractions, filtering becomes highly selective, leaving only the most mutually similar (hence homogeneous) neighbors, which can produce a slightly overfit but lower-error model on the test set.

VI. DISCUSSION

A. Scope and Applicability

SKETCHGUARD is presented as a wrapper around BALANCE for the purpose of theoretical instantiation and empirical comparison, but its design is not BALANCE-specific.

The only requirement for the underlying defense is that its filtering criterion is based on Euclidean distances between models—a property shared by UBAR [7], SCCLIP [8], and geometric-median-based methods [15]. Any such defense can be augmented with the sketch-exchange and selective-fetch phases of Algorithm 1, with the corresponding γ_{eff} adjustment to its filtering threshold.

The framework also naturally accommodates dynamic topologies. Because sketches are generated and exchanged independently each round, changes in the neighbor set between rounds require no additional coordination beyond what the base DFL protocol already handles. Our ER topology experiments, which resample the graph each round, confirm this.

B. Adaptive Attacks Against Sketch-Based Filtering

A natural concern is whether an adversary who knows SKETCHGUARD is deployed could craft a sketch-specific attack — for example, constructing a malicious model $\hat{\mathbf{w}}$ that lies just inside γ_{eff} in sketch space while being adversarially directed in full-precision space. We show analytically that the damage such an attack can cause is tightly bounded, and quantify that bound.

The acceptance region in full-precision space is bounded. By Lemma IV.8, any model $\hat{\mathbf{w}}_j$ that passes the sketch filter at round t satisfies:

$$\|\hat{\mathbf{w}}_j - \mathbf{w}_i^{t+1/2}\| \leq \gamma_{\text{eff}} \exp(-\kappa t/T) \|\mathbf{w}_i^{t+1/2}\| \leq \gamma_{\text{eff}} \psi. \quad (16)$$

This bound holds regardless of the attacker's strategy or intent. An adversary that optimally positions $\hat{\mathbf{w}}_j$ to be adversarially directed while remaining within the sketch acceptance region is still constrained to lie within a ball of radius $\gamma_{\text{eff}}\psi$ around $\mathbf{w}_i^{t+1/2}$ in full-precision space. The worst-case contribution of such a model to the aggregation is therefore bounded by $\gamma_{\text{eff}}\psi(1-\alpha)$, which is exactly the third term appearing in Theorems IV.9 and IV.10. This is not qualitatively different from the residual risk in full-precision filtering with threshold γ_{eff} — the two settings have identical worst-case aggregation error bounds.

The gap over full-precision filtering is small and quantified. The sketch-specific attack surface relative to full-precision filtering is precisely the $\sqrt{(1+\epsilon)/(1-\epsilon)}$ expansion of the acceptance radius. For $\epsilon = 0.2$, this is a factor of ≈ 1.22, meaning the attacker can place $\hat{\mathbf{w}}_j$ at most 22% further from $\mathbf{w}_i^{t+1/2}$ than would be possible under full-precision filtering. Whether this marginal expansion can be exploited depends on the geometry of the honest model distribution relative to the acceptance boundary — and our sensitivity experiments (Section V-E), which show stable TER across compression ratios up to 13,000:1, confirm empirically that honest neighbors are well inside the acceptance region and the 22% expansion does not admit additional Byzantine models.

The verification step closes the remaining protocol-level gap. An adversary might attempt to pass Phase 2 with a carefully crafted sketch and then submit a different full model in Phase 4. The sketch recomputation in Line 9 of Algorithm 1 detects any such mismatch deterministically, since Count Sketch is a deterministic function of the model for fixed hash functions. This attack vector is therefore closed unconditionally, not probabilistically.

C. Limitations

Bounded parameter norms. The convergence proofs rely on Assumption IV.4, which bounds both model norms ($\|\mathbf{w}_i\| \leq \psi$) and gradient norms ($\|\nabla F(\mathbf{w}_i)\| \leq \rho$). This is standard in the Byzantine DFL literature [2], [8] and can be enforced in practice via gradient clipping or projected gradient descent. However, without explicit clipping, these bounds may be violated in early training rounds when models are far from convergence, potentially loosening the theoretical guarantees before the training trajectory stabilises. In practice, our experiments show stable behaviour without explicit clipping, suggesting the bounds hold implicitly for the datasets and architectures evaluated, but this cannot be guaranteed for arbitrary models.

Shared hash functions. The distance-preservation property of Count Sketch (Lemma II.1) requires all clients to use identical hash and sign functions, so that sketch-domain distances are comparable across clients. This is trivially satisfied by seeding the hash functions with the shared model dimension d or a globally known seed agreed upon before training. In practice, this is a one-time coordination step with negligible cost — far cheaper than any round of model exchange. However, in fully asynchronous or open-membership DFL systems where clients join dynamically, ensuring hash function consistency requires a lightweight protocol (e.g., publishing the seed alongside the model architecture specification), which adds a small overhead not accounted for in our communication analysis.

Synchronous round structure. The analysis and experiments assume a synchronous protocol: in each round, all clients train locally, exchange sketches, and aggregate before the next round begins. This is consistent with the literature we evaluated against but does not capture asynchronous DFL settings where clients operate at different speeds or where stragglers delay aggregation. In asynchronous settings, a client may receive sketches from neighbours that were computed at different training rounds, making the sketch-domain distance comparisons less meaningful. Extending SKETCHGUARD to asynchronous DFL, where staleness must be accounted for in both the filtering threshold and the sketch validity window, is a natural direction for future work.

VII. CONCLUSIONS

We proposed SKETCHGUARD, a framework that removes the fundamental coupling between Byzantine filtering and full-model communication in decentralized federated learning. By exchanging compact Count Sketches for neighbor screening and fetching full models only from accepted neighbors, SKETCHGUARD reduces per-round communication from $O(d|\mathcal{N}_i|)$ to $O(k|\mathcal{N}_i| + d|\mathcal{S}_i^t|)$, with a sketch size k that is independent of model dimension d.

Our theoretical analysis establishes that Count Sketch's distance-preservation guarantee translates directly into a bounded degradation in filtering quality: sketch-based filtering is equivalent to full-precision filtering with a $\sqrt{(1+\epsilon)/(1-\epsilon)}$ inflated threshold, and convergence rates in both strongly convex and non-convex settings are preserved with no structural change. Experiments across three federated benchmarks, five network topologies, and four attack types confirm that SKETCHGUARD matches state-of-the-art robustness within 0.5 percentage points of TER while reducing computation by up to 82% and communication by 50–70% under adversarial conditions (with negligible overhead in benign settings). Robustness is stable across compression ratios up to 13,000:1, making the method robust to aggressive compression choices that may be forced by hardware constraints in practice.

ACKNOWLEDGMENT

This work is supported by the Australian Research Council (ARC) through Discovery Project grant DP240102088.

CODE AVAILABILITY

The SKETCHGUARD source code and all experiment artifacts are available at https://doi.org/10.5281/zenodo.17223405.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282.

[2] M. Fang, Z. Zhang, Hairi, P. Khanduri, J. Liu, S. Lu, Y. Liu, and N. Gong, "Byzantine-robust decentralized federated learning," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '24. Association for Computing Machinery, 2024, p. 2874–2888.

[3] E. T. Martínez Beltrán, M. Q. Pérez, P. M. S. Sánchez, S. L. Bernal, G. Bovet, M. G. Pérez, G. M. Pérez, and A. H. Celdrán, "Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2983–3013, 2023.

[4] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[5] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[6] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[7] S. Guo, T. Zhang, H. Yu, X. Xie, L. Ma, T. Xiang, and Y. Liu, "Byzantine-resilient decentralized stochastic gradient descent," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 4096–4106, 2022.

[8] L. He, S. P. Karimireddy, and M. Jaggi, "Byzantine-robust decentralized learning via clippedgossip," *arXiv preprint arXiv:2202.01545*, 2022.

[9] D. Cajaraville-Aboy, A. Fernández-Vilas, R. P. Díaz-Redondo, and M. Fernández-Veiga, "Byzantine-robust aggregation for securing decentralized federated learning," *IEEE Access*, vol. 13, pp. 190 947–190 963, 2025.

[10] E. M. El-Mhamdi, S. Farhadkhani, R. Guerraoui, A. Guirguis, L.-N. Hoang, and S. Rouault, "Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning)," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 25 044–25 057.

[11] P. Kairouz and H. B. McMahan, "Advances and open problems in federated learning," *Foundations and trends in machine learning*, vol. 14, no. 1-2, pp. 1–210, 2021.

[12] M. Rangwala and R. Buyya, "Trustmesh: A blockchain-enabled trusted distributed computing framework for open heterogeneous iot environments," in *2025 IEEE 22nd International Conference on Software Architecture (ICSA)*. IEEE, 2025, pp. 131–141.

[13] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," in *Automata, Languages and Programming*, P. Widmayer, S. Eidenbenz, F. Triguero, R. Morales, R. Conejo, and M. Hennessy, Eds. Springer Berlin Heidelberg, 2002, pp. 693–703.

[14] P. Sun, X. Liu, Z. Wang, and B. Liu, "Byzantine-robust decentralized federated learning via dual-domain clustering and trust bootstrapping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 24 756–24 765.

[15] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1142–1154, 2022.

[16] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to Byzantine-Robust federated learning," in *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, Aug. 2020, pp. 1605–1622.

[17] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 108. PMLR, 26–28 Aug 2020, pp. 2938–2948.

[18] V. Shejwalkar and A. Houmansadr, "Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning," in *NDSS*, 2021.

[19] C. Xie, O. Koyejo, and I. Gupta, "Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation," in *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, ser. Proceedings of Machine Learning Research, vol. 115. PMLR, 22–25 Jul 2020, pp. 261–270.

[20] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[21] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified sgd with memory," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[22] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-i.i.d. data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3400–3413, 2020.

[23] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[24] F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi, "Federated learning with compression: Unified analysis and sharp guarantees," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 130. PMLR, 13–15 Apr 2021, pp. 2350–2358.

[25] E. Gorbunov, S. Horváth, P. Richtárik, and G. Gidel, "Variance reduction is an antidote to byzantines: Better rates, weaker assumptions and communication compression as a cherry on the top," in *The Eleventh International Conference on Learning Representations*, 2023.

[26] A. Rammal, K. Gruntkowska, N. Fedin, E. Gorbunov, and P. Richtarik, "Communication compression for Byzantine robust learning: New efficient algorithms and improved rates," in *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 238. PMLR, 02–04 May 2024, pp. 1207–1215.

[27] D. Rothchild, A. Panda, E. Ullah, N. Ivkin, I. Stoica, V. Braverman, J. Gonzalez, and R. Arora, "FetchSGD: Communication-efficient federated learning with sketching," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 13–18 Jul 2020, pp. 8253–8265.

[28] G. Garrigos and R. M. Gower, "Handbook of convergence theorems for (stochastic) gradient methods," *arXiv preprint arXiv:2301.11235*, 2023.

[29] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar, "LEAF: A benchmark for federated settings," in *Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019.