

Byzantine-Resilient Collaborative Learning in Heterogeneous Resource-Adaptive Systems

Yuan Yuan, *Member, IEEE*, Lin Sun, Yan Li, *Student Member, IEEE*, Xiao Zhang *Member, IEEE*, Yifei Zou, *Member, IEEE*, Falko Dressler, *Fellow, IEEE*, Dongxiao Yu, *Senior Member, IEEE*

Abstract—Collaborative learning across heterogeneous resource-adaptive systems has emerged as a promising paradigm for enabling large-scale edge intelligence. However, the heterogeneity of resources and model components often leads to inconsistent optimization behaviors among participants, which makes the learning process highly sensitive to unreliable or even malicious nodes. To tackle this challenge, we propose a novel Byzantine Resilient Collaborative Learning algorithm (BRCL), which achieves Byzantine robustness while adapting to the dynamic and resource-constrained nature of heterogeneous environments. The proposed BRCL framework integrates adaptive resource scheduling with robust aggregation to maintain learning efficiency under adversarial conditions. We theoretically prove that BRCL attains an asymptotically optimal convergence rate of $O(1/\sqrt{TK}(\Gamma^* - f))$, where Γ^* is the minimum covering number, and f is the number of Byzantine clients. Special cases show that BRCL generalizes several classical algorithms, including FedAvg and OAP, under different system configurations. Extensive experiments on multiple learning tasks demonstrate that BRCL consistently outperforms state-of-the-art methods, achieving at least a 4.2% improvement in accuracy on average, while preserving robustness and resource efficiency in heterogeneous edge systems.

Index Terms—Byzantine resilient, collaborative learning, resource-adaptive, heterogeneity, edge computing.

I. INTRODUCTION

COLLABORATIVE learning enables distributed edge devices to jointly train machine learning models without directly sharing their local data, and has become an important paradigm for large-scale edge intelligence [18], [19], [25]. However, practical edge systems face two fundamental challenges. First, edge devices are typically heterogeneous and resource-constrained in terms of computation, memory, and communication [8], [16], [20]. Second, distributed learning is vulnerable to unreliable or malicious devices, whose arbitrary behaviors can be modeled as Byzantine faults [6], [17]. These two challenges become particularly intertwined when resource-constrained clients are allowed to train only part of the global model.

Y. Yuan is with the School of Artificial Intelligence, Shandong University, Jinan, 250000, China. E-mail: yyuan@sdu.edu.cn

L. Sun is with the China Unicom Research Institute of Data Science and Artificial Intelligence, Beijing, 100013, China. E-mail: sunl139@chinaunicom.cn

Y. Li, X. Zhang, and Y. Zou are with the School of Computer Science and Technology, Shandong University, Qingdao, 266237, China. E-mail: {xiaozhang, yfzou}@sdu.edu.cn; yhyanli@mail.sdu.edu.cn

F. Dressler is with the School of Electrical Engineering and Computer Science, TU Berlin, Berlin, 10587, Germany. E-mail: dressler@ccs-labs.org.

D. Yu is with the School of Cryptologic Science and Engineering, State Key Laboratory of Cryptography and Digital Economy Security, Shandong University, Jinan, Shandong, 250101, China. E-mail: dxyu@sdu.edu.cn

Manuscript received MM DD, YYYY; revised MM DD, YYYY.

Resource-adaptive partial-model training provides an effective way to accommodate heterogeneous devices by allowing clients to train sub-models according to their available resources [24]. However, partial-model training fundamentally changes the structure of Byzantine robustness. Since different clients may maintain different parts of the global model, each model region can be updated by a different subset of clients. Consequently, Byzantine robustness is no longer determined solely by the global client population. In particular, a benign majority among all participating clients does not necessarily guarantee sufficient benign participation in every model region. This makes Byzantine robustness inherently region-dependent.

Existing Byzantine-resilient learning methods mainly mitigate malicious updates through robust aggregation. Representative approaches include Multi-Krum [1], Trimmed Mean [22], Bulyan [11], and filter-based aggregation methods such as AFA [10]. These methods are primarily designed for settings where clients optimize a common set of model parameters. Other studies consider dimensionality reduction or gradient splitting to improve Byzantine robustness [9], [15]. However, they do not explicitly account for resource-dependent partial-model participation, where both the participating clients and the Byzantine composition may vary across model regions.

More importantly, resource adaptation and Byzantine robustness are intrinsically coupled in such systems. More aggressive pruning reduces the computation and communication burden of resource-constrained clients, but may also decrease the number of clients covering certain model regions, thereby reducing the redundancy available for robust aggregation. Therefore, the key challenge is not only to defend against Byzantine updates, but also to maintain sufficient regional coverage under heterogeneous resource constraints. This raises a fundamental question: *how can resource-adaptive collaborative learning achieve Byzantine robustness when different model regions are trained by different subsets of clients?*

To address this challenge, we propose BRCL, a Byzantine-Resilient Collaborative Learning framework for resource-heterogeneous edge systems. BRCL integrates resource-adaptive sub-model training with region-wise Byzantine-resilient aggregation. According to client resources, the global model is adaptively partitioned into sub-models for local training. For each model region, the server constructs a robust reference using the geometric median, filters suspicious updates, and aggregates the retained updates. In this way, robust aggregation is performed according to the actual participants of each model region, enabling BRCL to jointly accommodate resource heterogeneity and Byzantine behaviors.

The main contributions of this work are summarized as follows:

- We propose **BRCL**, a Byzantine-resilient collaborative learning framework for resource-heterogeneous edge systems. **BRCL** integrates resource-adaptive sub-model training with region-wise robust filtering and aggregation to address the region-dependent robustness caused by partial-model participation.
- We introduce the minimum regional coverage Γ^* and establish convergence guarantees that characterize the effects of regional coverage and Byzantine clients. Under the stated conditions, the optimization term scales as $O(1/\sqrt{TK(\Gamma^* - f)})$.
- We conduct extensive experiments across multiple datasets, model architectures, attack types, and Byzantine ratios. The results demonstrate the effectiveness and efficiency of **BRCL** compared with representative resource-adaptive and Byzantine-resilient methods.

II. RELATED WORK

A. Resource-constrained Collaborative Learning

Various methods have been proposed to cope with network resource limitations, aiming to minimize communication and computation. These methods include quantization [4], [14], sparsification [3], [21], knowledge distillation [7], and model pruning [2], [5], [23]. To address the challenges of communication bottleneck and scalability in federated learning, FedPAQ [14] proposes periodic averaging and quantization. Integrating federated learning with top-k sparse methods, the study in [3] achieves efficient communication. PruneFL [5] employs structural pruning to reduce model size and computational cost by eliminating redundant connections. HeteroFL [2] introduces an adaptive algorithm that assigns model subsets with narrower layers to clients based on their computational capabilities. Similar to model pruning, its goal is to optimize resource utilization. While these approaches address the resource constraints of edge networks, they lack guarantees for Byzantine robustness. Our work, in contrast, considers both resource constraints and Byzantine robustness.

B. Byzantine Robust Collaborative Learning

In recent years, the security of distributed learning has attracted increasing attention, among which the Byzantine threat is one of the most critical challenges. Blanchard et al. [1] proved that Multi-Krum ensures convergence to a stationary point for non-convex objectives when the number of Byzantine workers is bounded and standard smoothness assumptions hold. Yin et al. [22] further showed that Byzantine-robust gradient descent achieves convergence rates comparable to SGD up to an additional bias induced by Byzantine workers. Coordinate-wise methods such as Trimmed Mean [22] achieve optimal statistical error rates for strongly convex objectives, with estimation error scaling as $O(f/n)$ under bounded variance assumptions. Bulyan [11] further improves robustness in high-dimensional settings by combining Multi-Krum and trimmed mean. Geometric median-based approaches [13] provide robustness to up to 50% Byzantine workers and guarantee

convergence with error depending on client heterogeneity. In contrast, methods such as FABA and AFA [12], rely on heuristic filtering and generally lack rigorous convergence guarantees. Liu et al. [9] established convergence guarantees for gradient splitting under heterogeneous data, showing convergence to a neighborhood of the optimal solution with an error depending on the splitting strategy and Byzantine ratio. However, their approach assumes uniform model partitioning. In contrast, our method considers resource-adaptive sub-model training and establishes convergence guarantees under Byzantine settings in this more general scenario, explicitly characterizing the impact of both model adaptivity and adversarial behavior.

III. PRELIMINARIES

A. System Model

We consider a collaborative edge-learning system with one central server and N clients, indexed by $n \in [N] := \{1, \dots, N\}$, as illustrated in Fig. 1. Training proceeds for K communication rounds. At round k , the server maintains the global model $x_k \in \mathbb{R}^d$ and assigns a resource-adaptive sub-model to each client. Client n holds a local dataset D_n and is characterized by a computation budget $c_n > 0$ and a communication budget $b_n > 0$. The collection $\{(c_n, b_n)\}_{n=1}^N$ captures system heterogeneity.

Because low-resource clients may be unable to train and communicate the complete model, we use structured partial-model training. The server partitions the model into disjoint regions and assigns a subset of complete regions to each client.

B. Resource-Aware Model Partition and Pruning

The global coordinates are partitioned into I disjoint regions $\{\mathcal{I}_i\}_{i=1}^I$:

$$\bigcup_{i=1}^I \mathcal{I}_i = \{1, \dots, d\}, \quad \mathcal{I}_i \cap \mathcal{I}_j = \emptyset, \quad i \neq j. \quad (1)$$

For $v \in \mathbb{R}^d$, v^i denotes its restriction to $\mathcal{I}_i$, $d_i := |\mathcal{I}_i|$, and $r^i \in \{0, 1\}^d$ denotes the indicator vector of region $\mathcal{I}_i$.

Define normalized resource levels

$$\hat{c}_n = \min\left(1, \frac{c_n}{C_{\text{tgt}}}\right), \quad \hat{b}_n = \min\left(1, \frac{b_n}{B_{\text{tgt}}}\right), \quad (2)$$

the fused resource score

$$R_n = a\hat{c}_n + (1-a)\hat{b}_n, \quad a \in [0, 1], \quad (3)$$

and the retention ratio

$$k_n^* = \min\left\{\max(R_n^p, k_{\min}), \frac{b_n}{M_0 q}\right\}. \quad (4)$$

Here, $p > 0$ controls sensitivity to available resources, $k_{\min} > 0$ is the minimum retention ratio, M_0 is the base-model size, and q is the parameter-representation factor.

At round k , $z_{k,n}^i \in \{0, 1\}$ indicates whether region i is assigned to client n . The mask is

$$m_{k,n} = \sum_{i=1}^I z_{k,n}^i r^i, \quad (5)$$

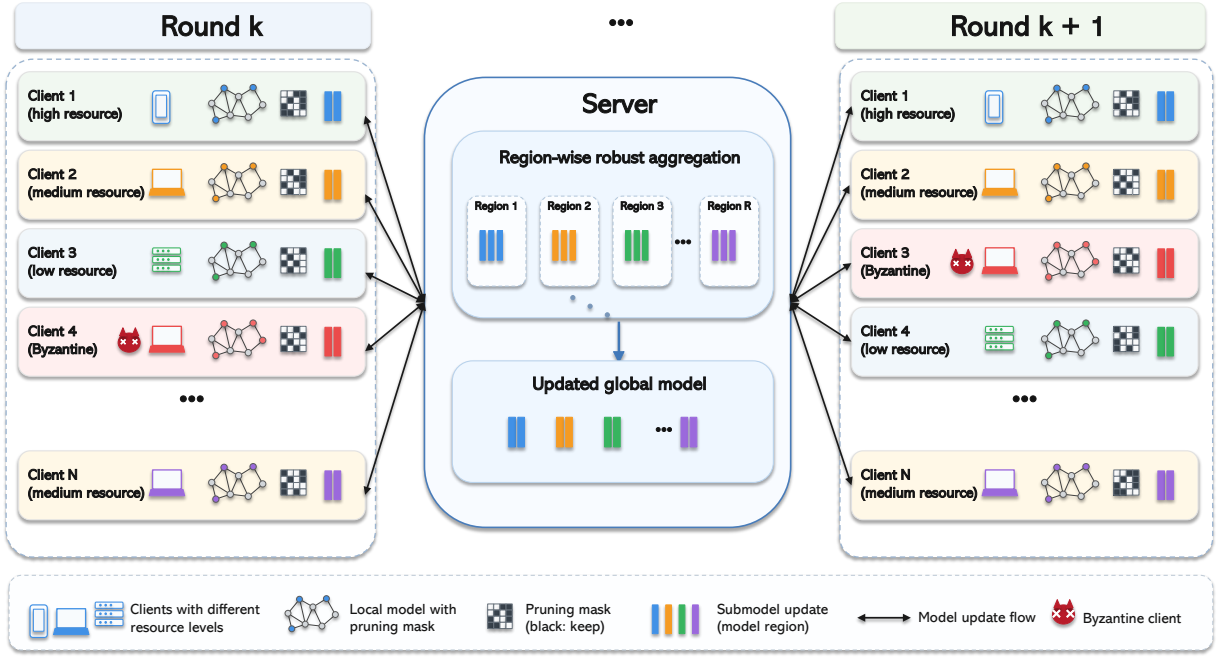


Fig. 1. An overview of the edge learning system model, where multiple edge devices train their own sub-models using their local data set and collaboratively build the global model under the coordination of a central server. The central server repeatedly aggregates local sub-models and broadcasts the global models to the client round by round. The clients are vulnerable to malicious attacks so that some of them become Byzantine clients and the set of Byzantine nodes can change during the training process (e.g. client 4 at round k and then client 3 at round $k + 1$).

subject to

$$\frac{\|m_{k,n}\|_0}{d} \leq k_n^*. \quad (6)$$

Thus, each selected region is trained and communicated as a complete block, keeping the client-side pruning rule consistent with server-side regional aggregation.

C. Byzantine Setting and Regional Coverage

The Byzantine population may vary across rounds. Let

$$\mathcal{M}_k \subseteq [N], \quad |\mathcal{M}_k| \leq f < \frac{N}{2}, \quad (7)$$

be the unknown true Byzantine set at round k . A Byzantine client may arbitrarily deviate from local training and transmit an arbitrary update. $\mathcal{M}_k$ is used only in the analysis.

For region i , define

$$\mathcal{N}_k^i := \{n \in [N] : z_{k,n}^i = 1\}, \quad (8)$$

and its latent Byzantine and benign subsets

$$\mathcal{M}_k^i := \mathcal{M}_k \cap \mathcal{N}_k^i, \quad \mathcal{H}_k^i := \mathcal{N}_k^i \setminus \mathcal{M}_k^i, \quad (9)$$

with $f_k^i := |\mathcal{M}_k^i|$. The exact regional benign-majority requirement is

$$f_k^i < \frac{|\mathcal{N}_k^i|}{2}, \quad \forall k, i. \quad (10)$$

Define

$$\Gamma^* := \min_{k,i} |\mathcal{N}_k^i|. \quad (11)$$

Since $f_k^i \leq f$, the sufficient global condition is

$$\Gamma^* \geq 2f + 1. \quad (12)$$

Operationally, the server can monitor

$$\Gamma_k := \min_i |\mathcal{N}_k^i|. \quad (13)$$

When coverage approaches the robustness boundary, the server can reduce pruning aggressiveness or assign the under-covered region to additional available clients in the next mask-generation step.

D. Optimization Objective

Client n has local objective

$$F_n(x) := \frac{1}{|D_n|} \sum_{\xi \in D_n} f(x; \xi). \quad (14)$$

The global objective is

$$\min_{x \in \mathbb{R}^d} F(x), \quad F(x) := \frac{1}{N} \sum_{n=1}^N F_n(x). \quad (15)$$

The local data distributions need not be identical; statistical heterogeneity is modeled in Assumption V.5. The goal is to minimize F under heterogeneous resource budgets and unknown, round-varying Byzantine sets.

IV. ALGORITHM

BRCL follows four consistent stages in each communication round: (i) resource-adaptive region assignment, (ii) normalized local training, (iii) region-wise robust detection and aggregation, and (iv) global-model reconstruction. Here, η denotes the learning rate, τ the filtering threshold, $z_{k,n}^i$ the region-assignment indicator, $m_{k,n}$ the mask, $G_{k,n}$ the normalized

cumulative update, $\mathcal{N}_k^i$ the regional participating set, $\widehat{\mathcal{B}}_k^i$ the detected-outlier set, and $\mathcal{S}_k^i$ the retained set. The latent true Byzantine sets $\mathcal{M}_k$ and $\mathcal{M}_k^i$ are used only for analysis.

A. Resource-Adaptive Mask Generation

Given the resource-dependent retention ratio in (4), the server selects complete regions for every client. Algorithm 1 shows the mask-generation procedure. The mask policy itself does not use Byzantine identities. After the masks are generated, the server can compute the realized regional coverages $|\mathcal{N}_k^i|$ directly from the assignment indicators.

Algorithm 1 Resource-Aware Mask Generation $\mathbb{P}$

- 1: **Input:** resources $\{(c_n, b_n)\}_{n=1}^N$; $(C_{\text{tgt}}, B_{\text{tgt}})$; $(a, p, k_{\min})$; model size M_0 ; factor q ; regions $\{\mathcal{I}_i\}_{i=1}^I$
 - 2: **for** $n = 1, \dots, N$ **do**
 - 3: $\hat{c}_n = \min(1, c_n/C_{\text{tgt}})$ and $\hat{b}_n = \min(1, b_n/B_{\text{tgt}})$
 - 4: $R_n = a\hat{c}_n + (1-a)\hat{b}_n$
 - 5: $k_n^* = \min\{\max(R_n^p, k_{\min}), b_n/(M_0q)\}$
 - 6: Set the parameter budget $d_n = \lfloor k_n^* d \rfloor$
 - 7: Select a subset of complete regions whose total size does not exceed d_n
 - 8: Set $z_{k,n}^i = 1$ for selected regions and 0 otherwise
 - 9: Construct $m_{k,n} = \sum_{i=1}^I z_{k,n}^i r^i$
 - 10: **end for**
 - 11: Compute $\mathcal{N}_k^i = \{n : z_{k,n}^i = 1\}$ for every region i
 - 12: **Output:** $\{m_{k,n}\}$, $\{z_{k,n}^i\}$, and $\{|\mathcal{N}_k^i|\}$
-

B. Normalized Local Update

At the beginning of round k , the server sends

$$x_{k,n,0} = x_k \odot m_{k,n} \quad (16)$$

to client n . A benign client performs T local SGD steps,

$$x_{k,n,t} = x_{k,n,t-1} - \eta g_n(x_{k,n,t-1}; \xi_{n,t-1}) \odot m_{k,n}. \quad (17)$$

After T steps, it forms the normalized cumulative update

$$G_{k,n} := \frac{x_{k,n,0} - x_{k,n,T}}{\eta} = \sum_{t=1}^T g_n(x_{k,n,t-1}; \xi_{n,t-1}) \odot m_{k,n}. \quad (18)$$

Thus, $G_{k,n}$ is an unscaled cumulative stochastic gradient, not a learning-rate-scaled model displacement. A Byzantine client $n \in \mathcal{M}_k$ may instead transmit an arbitrary vector.

C. Region-Wise Detection and Aggregation

For each region i , the server first computes the geometric median over all participating updates,

$$\tilde{G}_k^i := \text{GMed}\{G_{k,n}^i : n \in \mathcal{N}_k^i\} = \arg \min_{G \in \mathbb{R}^{d_i}} \sum_{n \in \mathcal{N}_k^i} \|G - G_{k,n}^i\|_2. \quad (19)$$

The server then defines the detected-outlier set

$$\widehat{\mathcal{B}}_k^i := \left\{ n \in \mathcal{N}_k^i : \|G_{k,n}^i - \tilde{G}_k^i\|_2 \geq \tau \|\tilde{G}_k^i\|_2 \right\}, \quad (20)$$

where $\tau > 0$ is the filtering threshold, and retains

$$\mathcal{S}_k^i := \mathcal{N}_k^i \setminus \widehat{\mathcal{B}}_k^i. \quad (21)$$

Crucially, $\widehat{\mathcal{B}}_k^i$ is a detector output and is not assumed to equal the true Byzantine set $\mathcal{M}_k^i$. To make this distinction explicit, define

$$\mathcal{E}_{k,\text{FP}}^i := \widehat{\mathcal{B}}_k^i \cap \mathcal{H}_k^i, \quad \mathcal{E}_{k,\text{FN}}^i := \mathcal{M}_k^i \cap \mathcal{S}_k^i. \quad (22)$$

The first set contains benign updates removed by mistake, while the second contains Byzantine updates that survive filtering.

If $\mathcal{S}_k^i \neq \emptyset$, BRCL uses

$$\bar{G}_k^i := \frac{1}{|\mathcal{S}_k^i|} \sum_{n \in \mathcal{S}_k^i} G_{k,n}^i. \quad (23)$$

If $\mathcal{S}_k^i = \emptyset$, it uses the fallback $\bar{G}_k^i = \tilde{G}_k^i$. Region i is then updated by

$$x_{k+1}^i = x_k^i - \eta \bar{G}_k^i. \quad (24)$$

Because the regions are disjoint, x_{k+1} is reconstructed by placing the updated blocks x_{k+1}^i back into their corresponding coordinates.

D. Complete BRCL Procedure

Algorithm 2 summarizes the complete workflow of BRCL. At each communication round, the server first generates resource-adaptive masks and distributes the corresponding sub-models to clients. After local training, the server collects the normalized updates and performs robust aggregation for each model region. Specifically, the geometric median $\tilde{G}_k^i$ is used as a robust reference to filter suspicious updates, and the retained updates in $\mathcal{S}_k^i$ are averaged to obtain $\bar{G}_k^i$. The resulting regional updates are then combined to update the global model for the next round.

V. ANALYSIS

We analyze exactly the update implemented in Algorithm 2. In particular, benign clients use the normalized cumulative update in (18), Byzantine identities are represented only by the latent sets $\mathcal{M}_k^i$, and the server aggregates the observable retained sets $\mathcal{S}_k^i$. Hence, the proof does not assume perfect outlier detection.

For a benign client n , define

$$D_{k,n} := \sum_{t=0}^{T-1} \mathbb{E} \|x_{k,n,t} - x_k\|^2, \quad (25)$$

and for region i define the reference update

$$\theta_k^i := T \nabla F^i(x_k). \quad (26)$$

A. Assumptions

Assumption V.1 (Smoothness). Each local objective F_n is differentiable and has L -Lipschitz gradients:

$$\|\nabla F_n(x) - \nabla F_n(y)\| \leq L\|x - y\|, \quad \forall x, y. \quad (27)$$

Assumption V.2 (Lower-bounded objective). There exists $F_{\inf} > -\infty$ such that

$$F(x) \geq F_{\inf}, \quad \forall x. \quad (28)$$

Algorithm 2 Byzantine-Resilient Collaborative Learning (BRCL)

```

1: Input:  $x_0, K, T$ , learning rate  $\eta$ , threshold  $\tau$ , mask policy  $\mathbb{P}$ 
2: for  $k = 0, \dots, K - 1$  do
3:   Generate  $\{m_{k,n}, z_{k,n}^i\}$  with Algorithm 1
4:   Send  $x_{k,n,0} = x_k \odot m_{k,n}$  and  $m_{k,n}$  to every client  $n$ 
5:   for  $n = 1, \dots, N$  do
6:     Receive transmitted update  $G_{k,n}$  from client  $n$ 
7:   end for
8:   for  $i = 1, \dots, I$  do
9:      $\mathcal{N}_k^i = \{n : z_{k,n}^i = 1\}$ 
10:     $\tilde{G}_k^i = \text{GMed}\{G_{k,n}^i : n \in \mathcal{N}_k^i\}$ 
11:     $\hat{\mathcal{B}}_k^i = \{n \in \mathcal{N}_k^i : \|G_{k,n}^i - \tilde{G}_k^i\| \geq \tau \|\tilde{G}_k^i\|\}$ 
12:     $\mathcal{S}_k^i = \mathcal{N}_k^i \setminus \hat{\mathcal{B}}_k^i$ 
13:    if  $\mathcal{S}_k^i \neq \emptyset$  then
14:       $\bar{G}_k^i = |\mathcal{S}_k^i|^{-1} \sum_{n \in \mathcal{S}_k^i} G_{k,n}^i$ 
15:    else
16:       $\bar{G}_k^i = \tilde{G}_k^i$ 
17:    end if
18:     $x_{k+1}^i = x_k^i - \eta \bar{G}_k^i$ 
19:  end for
20:  Reconstruct  $x_{k+1}$  from  $\{x_{k+1}^i\}_{i=1}^I$ 
21: end for
22: Benign ClientUpdate( $n, x_{k,n,0}, m_{k,n}$ ):
23: for  $t = 1, \dots, T$  do
24:   Sample  $\xi_{n,t-1}$  and update  $x_{k,n,t}$  according to (17)
25: end for
26: Return  $G_{k,n} = (x_{k,n,0} - x_{k,n,T})/\eta$ 
27: Byzantine client: may return an arbitrary  $G_{k,n}$ 

```

Assumption V.3 (Pruning-induced deviation). For every benign client n and round k ,

$$\|x_k - x_k \odot m_{k,n}\|^2 \leq \beta^2 \|x_k\|^2. \quad (29)$$

Assumption V.4 (Stochastic-gradient noise). For every benign client,

$$\mathbb{E}[g_{k,n,t} \mid x_{k,n,t}] = \nabla F_n(x_{k,n,t}), \quad (30)$$

and

$$\mathbb{E}[\|g_{k,n,t} - \nabla F_n(x_{k,n,t})\|^2 \mid x_{k,n,t}] \leq \sigma^2. \quad (31)$$

The stochastic-gradient noises form a martingale-difference sequence conditioned on the local iterates.

Assumption V.5 (Client heterogeneity). For every benign client n ,

$$\|\nabla F_n(x) - \nabla F(x)\|^2 \leq \delta^2, \quad \forall x. \quad (32)$$

B. Local-Update Error

Lemma V.6 (Accumulated local drift). Let

$$u := \eta LT, \quad q_\eta := \frac{1}{1 - 4u^2}. \quad (33)$$

If $u < 1/2$, then for every benign client,

$$D_{k,n} \leq q_\eta \left(2T\beta^2 \mathbb{E}\|x_k\|^2 + 2\eta^2 T^2 \sigma^2 \right)$$

$$+ \frac{16}{3} \eta^2 T^3 (\delta^2 + \mathbb{E}\|\nabla F(x_k)\|^2). \quad (34)$$

Proof. For a benign client, write $x_t = x_{k,n,t}$ and $x = x_k$. From the local recursion,

$$x_t - x = (x_{k,n,0} - x) - \eta \sum_{s=0}^{t-1} g_{k,n,s} \odot m_{k,n}. \quad (35)$$

Using Assumption V.3, the martingale-difference property in Assumption V.4, smoothness, and $\|\nabla F_n(x) - \nabla F(x)\|^2 \leq \delta^2$, we obtain

$$\begin{aligned} \mathbb{E}\|x_t - x\|^2 &\leq 2\beta^2 \mathbb{E}\|x\|^2 + 4\eta^2 t \sigma^2 \\ &\quad + 8\eta^2 t L^2 \sum_{s=0}^{t-1} \mathbb{E}\|x_s - x\|^2 \\ &\quad + 16\eta^2 t^2 (\delta^2 + \mathbb{E}\|\nabla F(x)\|^2). \end{aligned} \quad (36)$$

Summing over $t = 0, \dots, T - 1$ and using $\sum_{t=0}^{T-1} t \leq T^2/2$ and $\sum_{t=0}^{T-1} t^2 \leq T^3/3$ gives

$$\begin{aligned} D_{k,n} &\leq 2T\beta^2 \mathbb{E}\|x_k\|^2 + 2\eta^2 T^2 \sigma^2 \\ &\quad + 4\eta^2 L^2 T^2 D_{k,n} + \frac{16}{3} \eta^2 T^3 (\delta^2 + \mathbb{E}\|\nabla F(x_k)\|^2). \end{aligned} \quad (37)$$

Rearranging proves (34). $\square$

Lemma V.7 (Benign cumulative-update error). For any benign client $n \in \mathcal{H}_k^i$,

$$\begin{aligned} \mathbb{E}\|G_{k,n}^i - \theta_k^i\|^2 &\leq a_\sigma \sigma^2 + a_\delta \delta^2 \\ &\quad + a_\beta \beta^2 \mathbb{E}\|x_k\|^2 + a_g \mathbb{E}\|\nabla F(x_k)\|^2. \end{aligned} \quad (38)$$

where

$$\begin{aligned} a_\sigma &:= 3T + 6q_\eta \eta^2 L^2 T^3, \\ a_\beta &:= 6q_\eta L^2 T^2, \\ a_\delta &:= 3T^2 + 16q_\eta \eta^2 L^2 T^4, \\ a_g &:= 16q_\eta \eta^2 L^2 T^4. \end{aligned} \quad (39)$$

Proof. For a benign client, (18) gives

$$\begin{aligned} G_{k,n}^i - \theta_k^i &= \sum_{t=1}^T \left[g_n^i(x_{k,n,t-1}; \xi_{n,t-1}) - \nabla F_n^i(x_{k,n,t-1}) \right] \\ &\quad + \sum_{t=1}^T \left[\nabla F_n^i(x_{k,n,t-1}) - \nabla F_n^i(x_k) \right] \\ &\quad + T \left[\nabla F_n^i(x_k) - \nabla F^i(x_k) \right]. \end{aligned} \quad (40)$$

The three terms correspond to stochastic-gradient noise, local drift, and client heterogeneity, respectively. Applying $\|a + b + c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$, Assumptions V.1–V.5, and Lemma V.6 yields (38). $\square$

C. Region-Wise Byzantine Robustness

Let

$$n_k^i := |\mathcal{N}_k^i|, \quad h_k^i := |\mathcal{H}_k^i| = n_k^i - f_k^i. \quad (41)$$

Lemma V.8 (Geometric-median error under Byzantine updates). *Suppose that $f_k^i < n_k^i/2$. Then, for any reference vector θ ,*

$$\|\tilde{G}_k^i - \theta\| \leq \frac{2}{n_k^i - 2f_k^i} \sum_{n \in \mathcal{H}_k^i} \|G_{k,n}^i - \theta\|. \quad (42)$$

Consequently, with $\theta = \theta_k^i$,

$$\mathbb{E}\|\tilde{G}_k^i - \theta_k^i\|^2 \leq \kappa_{k,i}^2 \frac{1}{h_k^i} \sum_{n \in \mathcal{H}_k^i} \mathbb{E}\|G_{k,n}^i - \theta_k^i\|^2, \quad (43)$$

where

$$\kappa_{k,i} := \frac{2h_k^i}{n_k^i - 2f_k^i}. \quad (44)$$

Under (12),

$$\kappa_{k,i} \leq \kappa_* := \frac{2(\Gamma^* - f)}{\Gamma^* - 2f}. \quad (45)$$

Proof. The geometric median minimizes the sum of Euclidean distances, so

$$\sum_{n \in \mathcal{N}_k^i} \|\tilde{G}_k^i - G_{k,n}^i\| \leq \sum_{n \in \mathcal{N}_k^i} \|\theta - G_{k,n}^i\|. \quad (46)$$

For $n \in \mathcal{H}_k^i$,

$$\|\tilde{G}_k^i - G_{k,n}^i\| \geq \|\tilde{G}_k^i - \theta\| - \|G_{k,n}^i - \theta\|, \quad (47)$$

whereas for $n \in \mathcal{M}_k^i$,

$$\|\tilde{G}_k^i - G_{k,n}^i\| \geq \|\theta - G_{k,n}^i\| - \|\tilde{G}_k^i - \theta\|. \quad (48)$$

Substituting and canceling the Byzantine terms gives

$$(n_k^i - 2f_k^i)\|\tilde{G}_k^i - \theta\| \leq 2 \sum_{n \in \mathcal{H}_k^i} \|G_{k,n}^i - \theta\|, \quad (49)$$

which proves (42). Squaring and applying Cauchy–Schwarz yields (43). The uniform bound (45) follows from $n_k^i \geq \Gamma^*$ and $f_k^i \leq f$. $\square$

Combining Lemmas V.7 and V.8,

$$\mathbb{E}\|\tilde{G}_k^i - \theta_k^i\|^2 \leq \kappa_*^2 \left(a_\sigma \sigma^2 + a_\beta \beta^2 \mathbb{E}\|x_k\|^2 + a_\delta \delta^2 + a_g \mathbb{E}\|\nabla F(x_k)\|^2 \right). \quad (50)$$

Lemma V.9 (Filtered aggregate with imperfect detection). *For the aggregate in (23),*

$$\|\bar{G}_k^i - \tilde{G}_k^i\| \leq \tau \|\tilde{G}_k^i\|. \quad (51)$$

Consequently,

$$\mathbb{E}\|\bar{G}_k^i - \theta_k^i\|^2 \leq b_\sigma \sigma^2 + b_\beta \beta^2 \mathbb{E}\|x_k\|^2 + b_\delta \delta^2 + b_g \mathbb{E}\|\nabla F(x_k)\|^2, \quad (52)$$

where, with $c_\tau := 4\tau^2 + 2$,

$$\begin{aligned} b_\sigma &:= c_\tau \kappa_*^2 a_\sigma, \\ b_\beta &:= c_\tau \kappa_*^2 a_\beta, \\ b_\delta &:= c_\tau \kappa_*^2 a_\delta, \\ b_g &:= c_\tau \kappa_*^2 a_g + 4\tau^2 T^2. \end{aligned} \quad (53)$$

The bound does not require $\hat{\mathcal{B}}_k^i = \mathcal{M}_k^i$.

Proof. If $\mathcal{S}_k^i$ is nonempty, every retained update satisfies

$$\|G_{k,n}^i - \tilde{G}_k^i\| < \tau \|\tilde{G}_k^i\|, \quad n \in \mathcal{S}_k^i. \quad (54)$$

Therefore,

$$\begin{aligned} \|\bar{G}_k^i - \tilde{G}_k^i\| &\leq \frac{1}{|\mathcal{S}_k^i|} \sum_{n \in \mathcal{S}_k^i} \|G_{k,n}^i - \tilde{G}_k^i\| \\ &\leq \tau \|\tilde{G}_k^i\|. \end{aligned} \quad (55)$$

If $\mathcal{S}_k^i$ is empty, Algorithm 2 sets $\bar{G}_k^i = \tilde{G}_k^i$, so the same bound holds. Importantly, this argument applies to every retained update regardless of whether it is benign or an undetected Byzantine update. Thus, residual Byzantine updates in $\mathcal{E}_{k,\text{FN}}^i$ are explicitly allowed.

Next,

$$\begin{aligned} \|\bar{G}_k^i - \theta_k^i\|^2 &\leq 2\|\bar{G}_k^i - \tilde{G}_k^i\|^2 + 2\|\tilde{G}_k^i - \theta_k^i\|^2 \\ &\leq 2\tau^2 \|\tilde{G}_k^i\|^2 + 2\|\tilde{G}_k^i - \theta_k^i\|^2. \end{aligned} \quad (56)$$

Using

$$\|\tilde{G}_k^i\|^2 \leq 2\|\tilde{G}_k^i - \theta_k^i\|^2 + 2\|\theta_k^i\|^2 \quad (57)$$

and $\|\theta_k^i\|^2 = T^2 \|\nabla F^i(x_k)\|^2$, then substituting (50), gives (52). $\square$

False-positive detections remove some benign updates and can reduce the amount of useful information retained in the regional average, but they do not invalidate Lemma V.9. False-negative Byzantine updates may remain in $\mathcal{S}_k^i$, but any such update must satisfy the same threshold condition as every retained update. Hence, the revised analysis does not require perfect Byzantine identification.

VI. MAIN RESULT

Define

$$c_{\text{des}} := \frac{3}{4} - \eta LT - \frac{I(1 + \eta LT)b_g}{T^2}. \quad (58)$$

Theorem VI.1 (Convergence to a stationary neighborhood). *Suppose Assumptions V.1–V.5 hold, $\Gamma^* \geq 2f + 1$, $\eta LT < 1/2$, and $c_{\text{des}} > 0$. Let*

$$\Delta_F := F(x_0) - F_{\text{inf}}. \quad (59)$$

Then Algorithm 2 satisfies

$$\begin{aligned} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\|\nabla F(x_k)\|^2 &\leq \frac{\Delta_F}{c_{\text{des}} \eta T K} \\ &\quad + \frac{I(1 + \eta LT)}{c_{\text{des}} T^2} (b_\sigma \sigma^2 + b_\delta \delta^2) \\ &\quad + \frac{I(1 + \eta LT)b_\beta \beta^2}{c_{\text{des}} T^2} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\|x_k\|^2. \end{aligned} \quad (60)$$

Proof. By L -smoothness,

$$F(x_{k+1}) \leq F(x_k) + \langle \nabla F(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2. \quad (61)$$

Using $x_{k+1}^i - x_k^i = -\eta \bar{G}_k^i$, adding and subtracting $\theta_k^i = T \nabla F^i(x_k)$, and applying Young's inequality gives

$$\begin{aligned} \mathbb{E} \langle \nabla F(x_k), x_{k+1} - x_k \rangle &\leq -\frac{3}{4} \eta T \mathbb{E} \|\nabla F(x_k)\|^2 \\ &\quad + \frac{\eta}{T} \sum_{i=1}^I \mathbb{E} \|\bar{G}_k^i - \theta_k^i\|^2. \end{aligned} \quad (62)$$

Moreover,

$$\begin{aligned} \frac{L}{2} \mathbb{E} \|x_{k+1} - x_k\|^2 &\leq L \eta^2 T^2 \mathbb{E} \|\nabla F(x_k)\|^2 \\ &\quad + L \eta^2 \sum_{i=1}^I \mathbb{E} \|\bar{G}_k^i - \theta_k^i\|^2. \end{aligned} \quad (63)$$

Substituting Lemma V.9 and using (58) yields

$$\begin{aligned} \mathbb{E}[F(x_{k+1})] - \mathbb{E}[F(x_k)] &\leq -c_{\text{des}} \eta T \mathbb{E} \|\nabla F(x_k)\|^2 \\ &\quad + \frac{I \eta (1 + \eta LT)}{T} [b_\sigma \sigma^2 + b_\delta \delta^2] \\ &\quad + \frac{I \eta (1 + \eta LT)}{T} b_\beta \beta^2 \mathbb{E} \|x_k\|^2. \end{aligned} \quad (64)$$

Summing from $k = 0$ to $K - 1$ gives the correctly signed objective gap

$$\begin{aligned} c_{\text{des}} \eta T \sum_{k=0}^{K-1} \mathbb{E} \|\nabla F(x_k)\|^2 &\leq F(x_0) - \mathbb{E}[F(x_K)] \\ &\quad + \frac{I \eta (1 + \eta LT) K}{T} (b_\sigma \sigma^2 + b_\delta \delta^2) \\ &\quad + \frac{I \eta (1 + \eta LT)}{T} b_\beta \beta^2 \\ &\quad \times \sum_{k=0}^{K-1} \mathbb{E} \|x_k\|^2. \end{aligned} \quad (65)$$

Since $F(x_K) \geq F_{\text{inf}}$, dividing by $c_{\text{des}} \eta T K$ proves (60). $\square$

Corollary VI.2 (Explicit stepsize and convergence rate). *Let*

$$\eta_{K,T} := \min \left\{ \eta_0, \sqrt{\frac{\Gamma^* - f}{TK}} \right\}, \quad \eta_0 := \frac{1}{16LT}. \quad (66)$$

Suppose $\Gamma^ \geq 2f + 1$ and the filtering threshold is chosen such that*

$$I \left[\frac{16}{255} (4\tau^2 + 2)\kappa_*^2 + 4\tau^2 \right] \leq \frac{3}{16}, \quad \kappa_* := \frac{2(\Gamma^* - f)}{\Gamma^* - 2f}. \quad (67)$$

Then $\eta_{K,T} LT \leq 1/16$ and $c_{\text{des}} \geq 1/2$, so Theorem VI.1 gives

$$\begin{aligned} &\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla F(x_k)\|^2 \\ &\leq \frac{2\Delta_F}{\eta_{K,T} TK} + \frac{2I(1 + \eta_{K,T} LT)}{T^2} (b_\sigma \sigma^2 + b_\delta \delta^2) \\ &\quad + \frac{2I(1 + \eta_{K,T} LT) b_\beta \beta^2}{T^2} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|x_k\|^2. \end{aligned} \quad (68)$$

In particular, whenever

$$K \geq 256L^2T(\Gamma^* - f), \quad (69)$$

we have $\eta_{K,T} = \sqrt{(\Gamma^ - f)/(TK)}$, and hence*

$$\frac{2\Delta_F}{\eta_{K,T} TK} = \frac{2\Delta_F}{\sqrt{TK(\Gamma^* - f)}}. \quad (70)$$

Therefore, the optimization term in the convergence bound decays as

$$O\left(\frac{1}{\sqrt{TK(\Gamma^* - f)}}\right). \quad (71)$$

If, in addition, the residual terms in (68) vanish (e.g., in the idealized noiseless, homogeneous, and pruning-free case), the whole stationarity bound has the same rate. In the general stochastic and pruned setting, (68) establishes convergence to a stationary neighborhood rather than exact convergence to a stationary point.

Proof. For $\eta_{K,T} \leq \eta_0$, let $u = \eta_{K,T} LT \leq 1/16$. Then $q_\eta = (1 - 4u^2)^{-1} \leq 256/255$. From the definition of a_g in Lemma V.7,

$$\frac{a_g}{T^2} = 16q_\eta (\eta_{K,T} LT)^2 \leq \frac{16}{255}. \quad (72)$$

Consequently,

$$\frac{b_g}{T^2} \leq \frac{16}{255} (4\tau^2 + 2)\kappa_*^2 + 4\tau^2. \quad (73)$$

Since $\eta_{K,T} LT \leq 1/16$, condition (67) implies

$$c_{\text{des}} \geq \frac{3}{4} - \frac{1}{16} - \frac{3}{16} = \frac{1}{2}. \quad (74)$$

Substituting $c_{\text{des}} \geq 1/2$ into Theorem VI.1 proves (68). Finally, (69) is equivalent to $\sqrt{(\Gamma^* - f)/(TK)} \leq 1/(16LT)$, which proves (70). $\square$

Remark VI.3 (Interpretation of the explicit stepsize). The stepsize in (66) increases with the effective benign regional margin $\Gamma^* - f$ until the stability cap η_0 is reached. Thus, when the regional coverage is larger, the analysis permits a larger learning rate and yields the optimization-term scaling $O(1/\sqrt{TK(\Gamma^* - f)})$. The factor $\Gamma^* - f$ in this statement comes from the stepsize choice; it should not be interpreted as a variance-reduction factor of the geometric-median estimator. The latter would require additional statistical assumptions beyond Assumptions V.1–V.5.

Remark VI.4 (Role of regional coverage). Regional coverage affects Byzantine robustness through

$$\kappa_* = \frac{2(\Gamma^* - f)}{\Gamma^* - 2f}, \quad (75)$$

which becomes large as Γ^* approaches the geometric-median breakdown boundary $2f$. At the same time, the explicit stepsize in Corollary VI.2 uses the benign regional margin $\Gamma^* - f$. Hence, Γ^* has two distinct roles in the revised analysis: it controls the robustness constant through κ_* and enters the optimization-term scaling through the admissible stepsize. This distinction avoids attributing a $1/(\Gamma^* - f)$ variance-reduction effect to the geometric median without a supporting statistical assumption.

VII. EXPERIMENTS

We evaluate BRCL from four complementary perspectives. We first compare its Byzantine robustness with representative resource-adaptive, robust-aggregation, and federated-learning baselines on the original image- and text-classification tasks. We then extend the evaluation to deeper models and additional datasets, study the effect of regional coverage and the filtering threshold, and finally quantify the system overhead introduced by Byzantine-resilient regional aggregation.

A. Experimental Setup

1) *Datasets and Models*: We consider both image and text classification. For image classification, we use MNIST and CIFAR-10 with two-hidden-layer MLPs, denoted by MLP-MNIST and MLP-CIFAR10, respectively. The two models contain 159,010 and 1,233,210 parameters. For text classification, we use AGNews with FastText, denoted by FastText-AGNews, which contains 12,642,308 parameters. To examine whether the conclusions extend beyond shallow MLPs, we additionally evaluate ResNet-18 on CIFAR-10. The extended evaluation further includes FEMNIST and CelebA with ResNet-18.

2) *Data Partition and Training Configuration*: We use one server and $N = 10$ clients. The client datasets are generated using non-IID Dirichlet partitions with $\alpha = 0.15$ for MNIST, $\alpha = 1.5$ for CIFAR-10, and $\alpha = 0.5$ for AGNews. Unless otherwise specified, training lasts for $K = 300$ communication rounds, which is consistent with the curves reported in Figs. 2–4. Each benign client performs $T = 5$ local SGD steps per round. Momentum SGD is used with learning rate $\eta = 0.01$ for image classification and $\eta = 0.1$ for text classification. The experiments are executed on a server equipped with Intel(R) Xeon(R) Platinum 8358P 30-core processors and ten NVIDIA GeForce RTX 3090 GPUs.

3) *Resource Configuration*: The general formulation characterizes client resources by (c_n, b_n) and maps them to a model-retention ratio. In the controlled GPU experiments, we instantiate resource heterogeneity directly through a mixed pruning schedule rather than emulating device-specific physical values of (c_n, b_n) . Specifically, four clients use a 25% pruning rate, four use a 50% pruning rate, and two use a 75% pruning rate, corresponding to retention ratios of 75%, 50%, and 25%, respectively. Each partitioned layer tensor is divided into four blocks for pruning and assignment, while the embedding and fully connected layers are not further partitioned. This controlled setting makes the resource configuration reproducible while preserving heterogeneous local model capacities.

4) *Byzantine Attacks*: We consider Byzantine ratios of 10%, 20%, and 30%, corresponding to 1, 2, and 3 malicious clients among the ten clients. Malicious clients are randomly selected in each round, and attacks start from round 1. We evaluate three attacks:

- **Random attack**: the transmitted update is replaced by $G_{k,n} \leftarrow z$, where $z \sim \mathcal{N}(0, I)$ with scale 1.0.
- **Label-flipping attack**: malicious clients modify labels before local training; for the 10-class image tasks, $c \leftarrow 9 - c$.

TABLE I
MAIN EXPERIMENTAL CONFIGURATIONS.

Configuration	Value
Clients / communication rounds	10 / 300
Local steps	$T = 5$
Dirichlet α	MNIST: 0.15; CIFAR-10: 1.5; AGNews: 0.5
Pruning rates	25% (4), 50% (4), 75% (2)
Byzantine ratios	10%, 20%, 30%
Attack start	Round 1
Random attack	$G \leftarrow z$, $z \sim \mathcal{N}(0, I)$, scale 1.0
Safeguard attack	$G \leftarrow -0.6G$

- **Safeguard attack**: the malicious update is modified as $G_{k,n} \leftarrow -0.6G_{k,n}$.

Thus, the Random and Safeguard attacks manipulate the transmitted client update rather than directly modifying the global-model parameters.

B. Baselines

We compare BRCL with three groups of baselines. First, **IST**, **SplitFed**, and **OAP** represent heterogeneous or resource-adaptive collaborative-learning paradigms. Second, we combine representative robust aggregation rules, including **Geometric Median (GM)** and **Krum**, with applicable resource-adaptive baselines, yielding methods such as OAP+GM, OAP+Krum, SplitFed+GM, and SplitFed+Krum. Third, we include **FedMDFG** and **Ditto** as representative federated-learning baselines.

The original FedMDFG and Ditto formulations assume that clients maintain complete local models. To make the comparison compatible with the resource-constrained partial-model setting considered in this work, we construct block-wise adaptations. **Block-FedMDFG** partitions the model parameters into blocks and solves the multi-gradient problem for each block using the clients that contain that block before reconstructing the global update. **Block-Ditto** applies the same resource-dependent masks to the client-side personalized models, so that each client optimizes only its assigned partial model. We use these names throughout to distinguish the resource-constrained implementations from the original full-model algorithms.

C. Robustness under Byzantine Attacks

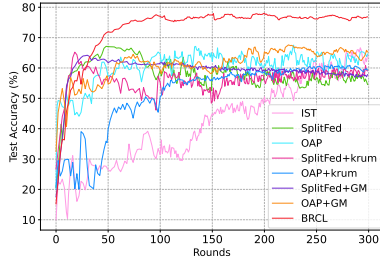
1) *Main Results on MNIST and CIFAR-10*: Table II compares the methods under three attack types and three Byzantine ratios. BRCL remains competitive across all evaluated settings and is particularly robust when the Byzantine ratio increases. The comparison with OAP+GM and OAP+Krum shows that simply attaching a conventional robust aggregator to a resource-adaptive method does not fully address the region-dependent robustness problem. Block-FedMDFG and Block-Ditto provide stronger federated-learning references under the same partial-model constraint, while BRCL explicitly performs robustness processing at the region level.

The results also show the expected effect of increasing the Byzantine ratio. As the ratio changes from 10% to 30%, the accuracy of most methods decreases because a larger fraction of the regional updates can be malicious. This trend is consistent with the role of f in the regional robustness condition.

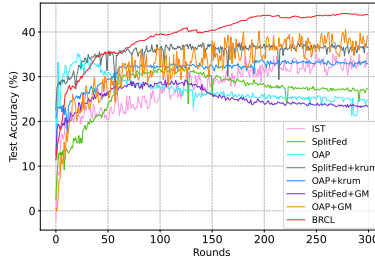
TABLE II

PERFORMANCE COMPARISON ON IMAGE CLASSIFICATION. THE LEFT PART REPORTS MLP-MNIST AND THE RIGHT PART REPORTS MLP-CIFAR10. THE VALUES DENOTE TEST ACCURACY (%). “–” INDICATES THAT THE CORRESPONDING RESULT IS UNAVAILABLE.

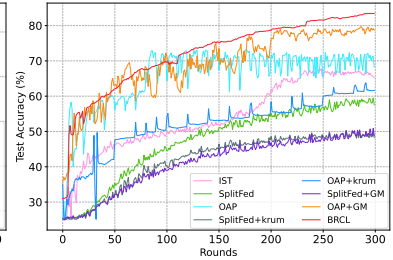
Methods	MNIST									CIFAR-10								
	Random			Label flipping			Safeguard			Random			Label flipping			Safeguard		
	30%	20%	10%	30%	20%	10%	30%	20%	10%	30%	20%	10%	30%	20%	10%	30%	20%	10%
IST	–	–	–	58.2	61.9	70.3	62.1	70.4	74.2	–	–	–	32.9	34.8	36.3	31.4	32.1	33.9
SplitFed	–	–	–	49.8	56.2	57.7	54.7	60.2	61.9	–	–	–	29.5	31.6	33.5	28.7	31.7	33.3
OAP	27.0	31.1	49.3	57.6	57.7	58.3	60.6	61.0	66.7	18.1	21.3	26.4	24.8	27.8	31.4	22.6	24.1	25.8
SplitFed+Krum	64.5	65.5	68.5	55.0	59.0	63.7	58.1	61.1	64.5	31.8	32.4	32.4	34.2	35.1	36.1	34.1	34.8	36.1
OAP+Krum	63.6	70.3	79.6	54.5	62.6	72.8	60.7	65.4	67.1	39.0	39.5	41.5	30.4	37.6	39.2	31.2	32.5	33.7
SplitFed+GM	67.7	70.1	72.0	65.6	66.1	67.1	57.4	70.1	68.5	31.7	33.5	35.5	29.1	30.6	32.1	25.8	26.1	26.7
OAP+GM	56.7	63.3	72.3	74.7	78.4	82.4	66.8	72.5	77.8	21.2	28.9	36.7	39.4	41.2	43.0	38.5	42.8	42.9
Block-FedMDFG	55.93	68.00	76.56	51.85	64.21	74.12	64.96	71.66	77.53	26.04	30.12	35.40	26.63	29.72	35.74	27.80	32.04	36.84
Block-Ditto	45.03	62.48	78.73	60.04	74.28	84.70	55.74	70.63	82.41	16.81	22.41	30.23	21.22	28.33	36.29	20.11	26.05	34.91
BRCL	78.7	81.2	83.2	81.5	82.8	83.4	78.3	80.8	82.5	41.9	44.3	46.1	41.8	42.6	45.1	39.8	43.1	44.9



(a) MLP-MNIST



(b) MLP-CIFAR10



(c) FastText-AGNews

Fig. 2. Training curves of different learning paradigms under the Safeguard attack.

TABLE III
PERFORMANCE COMPARISON ON FASTTEXT-AGNEWS.

Methods	Random			Label flipping			Safeguard		
	30%	20%	10%	30%	20%	10%	30%	20%	10%
IST	–	–	–	45.7	48.1	54.8	–	66.3	70.8
SplitFed	26.6	28.7	38.4	40.0	44.0	50.8	56.4	57.5	59.6
OAP	37.0	41.1	45.2	65.9	69.2	70.1	65.7	72.1	75.8
SplitFed+Krum	48.3	54.2	55.9	36.6	40.2	45.2	43.9	48.3	52.3
OAP+Krum	46.5	53.1	62.5	57.9	67.0	69.7	58.1	64.1	69.9
SplitFed+GM	50.8	51.0	53.3	44.7	48.9	53.3	47.4	48.8	52.5
OAP+GM	48.7	59.3	59.8	68.9	77.1	76.8	74.8	77.9	80.4
Block-FedMDFG	50.05	53.15	58.67	60.68	66.75	72.20	62.27	70.38	78.79
Block-Ditto	26.56	36.96	48.20	36.37	48.38	60.91	34.37	46.62	58.37
BRCL	51.2	62.3	66.2	84.3	85.0	87.1	77.9	83.4	84.9

2) *Text Classification on AGNews*: Table III shows a similar trend on AGNews. BRCL maintains strong performance across the three attacks and different Byzantine ratios. The relative severity of an attack is task dependent: for example, label flipping is less destructive than Random or Safeguard attacks for some learning paradigms, whereas other methods exhibit different sensitivities. This observation motivates evaluating multiple attack types rather than drawing conclusions from a single Byzantine strategy.

Figures 2–4 provide complementary views of the main results. Figure 2 compares representative learning paradigms, Fig. 3 illustrates the task-dependent effect of different attack types, and Fig. 4 shows the degradation caused by increasing the Byzantine ratio.

D. Evaluation with Deeper Models and Additional Datasets

We next examine whether the observed robustness extends beyond the original MLP/FastText settings.

As shown in Table IV, BRCL remains effective when the MLP is replaced by ResNet-18. Importantly, the expanded comparison includes not only OAP variants but also SplitFed-based robust methods, Block-FedMDFG, and Block-Ditto. This reduces the dependence of the conclusion on a single resource-adaptive baseline family.

Table V further extends the comparison to FEMNIST and CelebA. Under both Random and Safeguard attacks, BRCL achieves higher accuracy than the compared methods in the reported experiments. Together with the ResNet-18/CIFAR-10 results, these experiments indicate that the benefit of region-wise Byzantine-resilient aggregation is not restricted to the original datasets or shallow architectures.

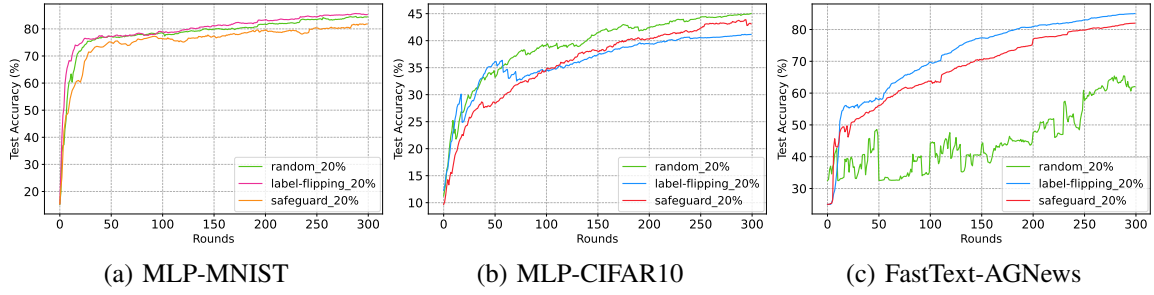


Fig. 3. Impact of different Byzantine attack types at a 20% Byzantine ratio.

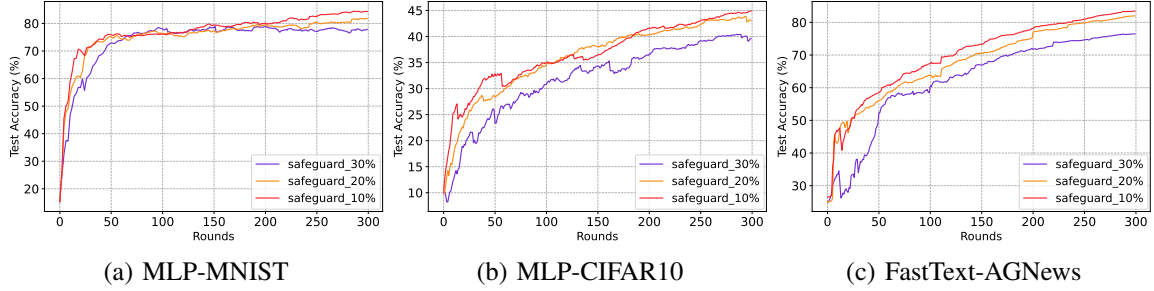


Fig. 4. Impact of the Byzantine-client ratio under the Safeguard attack.

TABLE IV
ACCURACY (%) ON CIFAR-10 WITH RESNET-18.

Method	Random		Safeguard	
	30%	10%	30%	10%
OAP	20.97	31.48	27.53	35.32
OAP+GM	35.26	42.62	31.61	37.04
OAP+Krum	36.05	47.71	52.00	51.28
SplitFed+GM	27.51	34.07	25.81	30.57
SplitFed+Krum	32.28	38.42	33.05	36.22
Block-FedMDFG	38.59	46.21	36.84	43.41
Block-Ditto	34.04	44.58	38.21	46.88
BRCL	51.79	54.94	55.60	57.31

TABLE V
ACCURACY (%) ON FEMNIST AND CELEBA WITH RESNET-18 UNDER A 20% BYZANTINE RATIO.

Method	FEMNIST		CelebA	
	Random	Safeguard	Random	Safeguard
OAP+GM	44.50	52.93	58.33	68.40
OAP+Krum	46.85	55.08	60.25	70.26
SplitFed+GM	34.24	40.27	50.54	58.17
SplitFed+Krum	38.59	45.12	54.07	62.09
Block-FedMDFG	42.08	50.38	55.77	66.55
Block-Ditto	38.48	50.77	58.12	70.12
BRCL	62.11	68.36	74.38	80.84

E. Impact of Regional Coverage and Filtering Threshold

1) *Regional Coverage*: The theoretical robustness condition depends on maintaining a benign majority within every model region. We therefore study the effect of the minimum regional coverage Γ^* . The additional experiments show substantial performance degradation when regional coverage approaches or falls below the Byzantine-tolerance boundary. In particular, the reported accuracies for $\Gamma^* = 2, 3$, and 4 are 12.9%, 26.5%, and 36.9%, respectively. These results are consistent with the role of the sufficient condition $\Gamma^* \geq 2f + 1$ in the analysis.

This observation also has a practical implication. At round k , the server can monitor $\Gamma_k = \min_i |\mathcal{N}_k^i|$. If resource-aware pruning causes a region to approach the required coverage boundary, the server can reduce pruning aggressiveness or assign that region to additional available clients in the subsequent mask-

generation step. Hence, regional coverage provides a direct operational link between resource adaptation and Byzantine tolerance.

2) *Filtering Threshold τ* : Figure 5 illustrates the impact of τ under label-flipping attacks with 30% Byzantine clients. A larger τ raises the rejection threshold and therefore makes filtering less aggressive. The observed behavior suggests a trade-off: a small threshold can reject more abnormal updates but may also remove benign updates, whereas a large threshold retains more benign information at the cost of allowing a larger residual deviation.

F. System Overhead

We finally quantify the additional system cost introduced by Byzantine-resilient regional aggregation. Table VI reports server-side aggregation time, mask-generation time, communication volume per round, and total training time.

As expected, BRCL introduces additional server-side computation because of region-wise geometric-median computation,

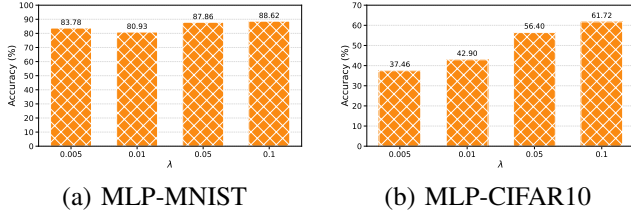
Fig. 5. Impact of the filtering threshold τ .

TABLE VI
SYSTEM OVERHEAD COMPARISON OF DIFFERENT METHODS.

Method	Agg. (ms/round)	Mask (ms/round)	Comm. (MB/round)	Training (min/round)
OAP	6.2	4.8	6.36	3.55
OAP+GM	25.7	4.8	6.36	3.64
OAP+Krum	18.5	4.8	6.36	3.61
SplitFed	6.3	—	492.56	5.90
SplitFed+GM	28.5	—	492.56	6.05
SplitFed+Krum	20.1	—	492.56	6.00
BRCL	56.2	6.3	6.36	3.84

outlier detection, and subsequent averaging. Its aggregation time is 56.2 ms/round, compared with 6.2 ms/round for OAP, while mask generation requires 6.3 ms/round compared with 4.8 ms/round for the OAP-based methods. Nevertheless, the effect on end-to-end training time is more limited: BRCL requires 3.84 min/round compared with 3.55 min/round for OAP. More importantly, the Byzantine-resilient operations are performed after the regional updates have been received and therefore do not require an additional client-server communication round. BRCL communicates 6.36 MB/round, the same volume as the OAP-based methods, whereas the SplitFed-based methods require 492.56 MB/round in the reported configuration.

BRCL introduces additional server-side computation for Byzantine resilience, while preserving the communication efficiency of resource-adaptive partial-model training.

VIII. CONCLUSION

In this paper, we proposed BRCL, a Byzantine-resilient collaborative learning framework for resource-heterogeneous edge systems. BRCL integrates resource-adaptive sub-model training with region-wise robust aggregation to defend against Byzantine updates. We established its convergence guarantee and demonstrated its effectiveness under different datasets, models, and Byzantine attacks. Future work will consider more dynamic and asynchronous edge-learning environments.

REFERENCES

- [1] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 119–129, 2017.
- [2] Enmao Diaio, Jie Ding, and Vahid Tarokh. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

- [3] Negar Foroutan Eghlidi and Martin Jaggi. Sparse communication for training deep networks. *CoRR*, abs/2009.09271, 2020.
- [4] Mahmudul Hasan, Yong Xiang, Md Palash Uddin, Jine Tang, John Yearwood, and Longxiang Gao. Communication-efficient federated learning via fractional-order gradient descent with adaptive momentum under non-iid data. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–16, 2026.
- [5] Yuang Jiang, Shiqiang Wang, Bong Jun Ko, Wei-Han Lee, and Leandros Tassioulas. Model pruning enables efficient federated learning on edge devices. *CoRR*, abs/1909.12326, 2019.
- [6] LESLIE LAMPORT, ROBERT SHOSTAK, and MARSHALL PEASE. The byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.
- [7] Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- [8] Junyu Liu, Chengyi Zhou, Min Sheng, Haojun Yang, Xinyu Huang, and Jiandong Li. Resource allocation for adaptive beam alignment in uav-assisted integrated sensing and communication networks. *IEEE J. Sel. Areas Commun.*, 43(1):350–363, 2025.
- [9] Yuchen Liu, Chen Chen, Lingjuan Lyu, Fangzhao Wu, Sai Wu, and Gang Chen. Byzantine-robust learning on heterogeneous data via gradient splitting. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 21404–21425. PMLR, 2023.
- [10] Vivek Mankotia, Ramesh Kumar Sunkaria, and Shashi Gurung. AFA: anti-flooding attack scheme against flooding attack in MANET. *Wirel. Pers. Commun.*, 130(2):1161–1190, 2023.
- [11] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3518–3527. PMLR, 2018.
- [12] Luis Muñoz-González, Kenneth T. Co, and Emil C. Lupu. Byzantine-robust federated machine learning through adaptive model averaging. *CoRR*, abs/1909.05125, 2019.
- [13] Krishna Pillutla, Sham M. Kakade, and Zaïd Harchaoui. Robust aggregation for federated learning. *IEEE Trans. Signal Process.*, 70:1142–1154, 2022.
- [14] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2021–2031. PMLR, 2020.
- [15] Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *28th Annual Network and Distributed System Security Symposium, NDSS 2021, virtually, February 21-25, 2021*. The Internet Society, 2021.
- [16] Youming Tao, Shuzhen Chen, Congwei Zhang, Di Wang, Dongxiao Yu, Xiuzhen Cheng, and Falko Dressler. Private over-the-air federated learning at band-limited edge. *IEEE Transactions on Mobile Computing*, 2024.
- [17] Youming Tao, Sijia Cui, Wenlu Xu, Haofei Yin, Dongxiao Yu, Weifa Liang, and Xiuzhen Cheng. Byzantine-resilient federated learning at edge. *IEEE Transactions on Computers*, 2023.
- [18] Haowei Wang, Jiayi Ji, Tianyu Guo, Yilong Yang, Xiaoshuai Sun, and Rongrong Ji. NICE: improving panoptic narrative detection and segmentation with cascading collaborative learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(10):8990–9004, 2025.
- [19] Junbo Wang, Amitangshu Pal, Qinglin Yang, Krishna Kant, Kaiming Zhu, and Song Guo. Collaborative machine learning: Schemes, robustness, and privacy. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [20] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. When edge meets learning: Adaptive control for resource-constrained distributed machine learning. In *IEEE INFOCOM 2018-IEEE conference on computer communications*, pages 63–71. IEEE, 2018.

- [21] Dezhong Yao, Wanning Pan, Yuexin Shi, Michael J. O'Neill, Yutong Dai, Yao Wan, Peilin Zhao, Hai Jin, and Limin Sun. Fedhm: Efficient federated learning for heterogeneous models via low-rank factorization. *Artif. Intell.*, 344:104333, 2025.
- [22] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter L. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5636–5645. PMLR, 2018.
- [23] Sifan Zhao, Tongtong Liu, Hai Jin, and Dezhong Yao. Spvit: Accelerate vision transformer inference on mobile devices via adaptive splitting and offloading. *IEEE Trans. Mob. Comput.*, 24(10):9303–9318, 2025.
- [24] Hanhan Zhou, Tian Lan, Guru Venkataramani, and Wenbo Ding. On the convergence of heterogeneous federated learning with arbitrary adaptive online model pruning. *CoRR*, abs/2201.11803, 2022.
- [25] Jie Zhou, Biao Jie, Zhengdong Wang, Zhixiang Zhang, Weixin Bian, Yang Yang, Hongwei Li, Fengyun Sun, and Mingxia Liu. Dcnnet: Double collaborative learning network on stationary-dynamic functional brain network for brain disease classification. *IEEE Trans. Image Process.*, 34:4026–4039, 2025.



Yuan Yuan received the BSc degrees from the School of Mathematical Sciences, Shanxi University in 2016, and the Ph.D. degree from the School of Computer Science and Technology, Shandong University, Qingdao, China, in 2021. She is currently an Associate Professor with School of Artificial Intelligence, Shandong University. Her research interests include distributed computing and distributed machine learning.



Lin Sun is currently the director of Innovative Product Development Department at China Unicom Research Institute of Data Science and Artificial Intelligence, Beijing, China. He received the B.S. and M.S. degrees from Northwestern Polytechnical University, Xi'an, China, in 2005 and 2008, respectively, and the Ph.D from Télécom SudParis and Université Pierre et Marie Curie, France, in 2013. His research interests include trustworthy collaborative computing, privacy computing, blockchain, and ubiquitous computing.



Yan Li is currently a PhD student in the School of Computer Science and Technology, Shandong University. He received her B.S. degree in Shandong University. His research interests include distributed systems and LLM inference optimization.



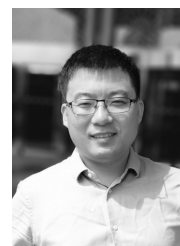
Xiao Zhang received his B.S. and Ph.D degree from Central South University and Nanjing University, China, respectively. He is now an associate professor in the School of Computer Science and Technology, Shandong University. Dr. Zhang's research interests include data mining, intelligent sensing, multi-task learning and federated learning.



Yifei Zou received the B.E. degree in computer science and technology from Computer School, Wuhan University, Wuhan, China, in 2016, and the Ph.D. degree in computer science from the Department of Computer Science, The University of Hong Kong, Hong Kong, in 2020. He is currently an Associate Professor with the School of Computer Science and Technology, Shandong University, Qingdao, China. His research interests include wireless networks, ad hoc networks, and distributed computing.



Falko Dressler received his M.Sc. and Ph.D. degrees from the Dept. of Computer Science, University of Erlangen in 1998 and 2003, respectively. He is a full professor and Chair for Data Communications and Networking at the School of Electrical Engineering and Computer Science, TU Berlin. Dr. Dressler has been associate editor-in-chief for IEEE Trans. on Mobile Computing and Elsevier Computer Communications as well as an editor for journals such as IEEE/ACM Trans. on Networking, IEEE Trans. on Network Science and Engineering, Elsevier Ad Hoc Networks, and Elsevier Nano Communication Networks. He has been chairing conferences such as IEEE INFOCOM, ACM MobiSys, ACM MobiHoc, IEEE VNC, IEEE GLOBECOM. He authored the textbooks Self-Organization in Sensor and Actor Networks published by Wiley & Sons and Vehicular Networking published by Cambridge University Press. He has been an IEEE Distinguished Lecturer as well as an ACM Distinguished Speaker. Dr. Dressler is an IEEE Fellow as well as an ACM Distinguished Member. He is a member of the German National Academy of Science and Engineering (acatech). He has been serving on the IEEE COMSOC Conference Council and the ACM SIGMOBILE Executive Committee. His research objectives include adaptive wireless networking (radio, visible light, molecular communications) and embedded system design (from microcontroller to Linux kernel) with applications in ad hoc and sensor networks, the Internet of Things, and cooperative autonomous driving systems.



Dongxiao Yu received the BSc degree in 2006 from the School of Mathematics, Shandong University and the PhD degree in 2014 from the Department of Computer Science, the University of Hong Kong. He became an associate professor in the School of Computer Science and Technology, Huazhong University of Science and Technology, in 2016. He is currently a professor in the School of Cryptologic Science and Engineering, State Key Laboratory of Cryptography and Digital Economy Security, Shandong University. His research interests include wireless networks, distributed computing and graph algorithms.