Certifiably Byzantine-Robust Federated Conformal Prediction

Mintong Kang [1] **Zhen Lin** [1] **Jimeng Sun** [1 2] **Cao Xiao** [3] **Bo Li** [1 4]

Abstract

Conformal prediction has shown impressive capacity in constructing statistically rigorous prediction sets for machine learning models with exchangeable data samples. The siloed datasets, coupled with the escalating privacy concerns related to local data sharing, have inspired recent innovations extending conformal prediction into federated environments with distributed data samples. However, this framework for distributed uncertainty quantification is susceptible to Byzantine failures. A minor subset of malicious clients can significantly compromise the practicality of coverage guarantees. To address this vulnerability, we introduce a novel framework Rob-FCP, which executes robust federated conformal prediction, effectively countering malicious clients capable of reporting arbitrary statistics in the conformal calibration process. We theoretically provide the conformal coverage bound of Rob-FCP in the Byzantine setting and show that the coverage of Rob-FCP is asymptotically close to the desired coverage level. We also propose a malicious client number estimator to tackle a more challenging setting where the number of malicious clients is unknown to the defender. We theoretically show the precision of the malicious client number estimator. Empirically, we demonstrate the robustness of Rob-FCP against various portions of malicious clients under multiple Byzantine attacks on five standard benchmark and real-world healthcare datasets.

1. Introduction

As deep neural networks (DNNs) achieved great success across multiple fields (He et al., 2016; Vaswani et al., 2017; Li et al., 2022b), quantifying the uncertainty of model predictions has become essential, especially in safety-conscious domains such as healthcare and medicine (Ahmad et al., 2018; Erickson et al., 2017; Kompa et al., 2021). For example, in sleep medicine domain, accurately classifying sleep stages (typically on EEG recordings) is crucial for understanding sleep disorders. Analogous to a human expert who may offer multiple possible interpretations of a single recording, it is desirable for a DNN to provide not just a singular prediction but a set of possible outcomes. In constructing such prediction sets, we often consider the following coverage guarantee: the prediction set should contain the true outcome with a pre-specified probability (e.g. 90%). *Conformal prediction* (Shafer & Vovk, 2008; Balasubramanian et al., 2014; Romano et al., 2020) demonstrates the capacity to provide such statistical guarantees for any black-box DNN with exchangeable data.

Meanwhile, the demand for training machine learning models on large-scale and diverse datasets necessitates model training across multiple sites and institutions. Federated learning (Konečný et al., 2016; Smith et al., 2017; McMahan et al., 2017; Bonawitz et al., 2019; Yang et al., 2019; Kairouz et al., 2021) offers an effective approach to collaboratively train a global model while preserving data privacy, as it enables training with distributed data samples without the requirement of sharing the raw data. For example, multiple hospitals ("clients") could jointly train a global clinical risk prediction model without sharing raw patient data. However, this introduces a unique challenge: the existence of malicious or negligent clients can negatively affect the training/testing of the global model.

Recently, federated conformal prediction (FCP) methods (Lu & Kalpathy-Cramer, 2021; Lu et al., 2023; Plassier et al., 2023; Humbert et al., 2023) provide rigorous bounds on the coverage rate with distributed data samples. However, FCP demonstrates vulnerability to *Byzantine failures* (Lamport et al., 2019), which are caused by uncontrollable behaviors of malicious clients. For example, a hospital's data could be corrupted with incorrect or even fabricated medical information due to human negligence or deliberate manipulation of data statistics (such as age, gender, or disease prevalence). In the Byzantine federated setting, the prediction coverage guarantees of FCP are broken, and the empirical marginal coverage is downgraded severely, even

[1]University of Illinois at Urbana-Champaign, USA [2]Carle's Illinois College of Medicine, USA [3]GE Healthcare, USA [4]University of Chicago, USA. Correspondence to: Mintong Kang <mintong2@illinois.edu>, Bo Li <lbo@illinois.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

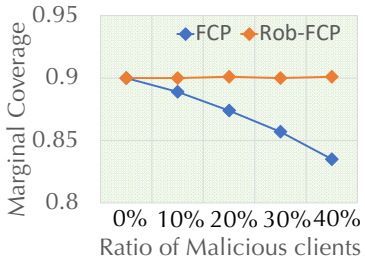


Figure 1: Coverage rate with different ratios of malicious clients on SHHS dataset. The desired coverage is 0.9.

with a small portion of malicious clients as Figure 1.

In this paper, we propose a robust federated conformal prediction algorithm, Rob-FCP, aimed at mitigating the impact of malicious clients on the coverage rate in Byzantine federated learning environments. The Rob-FCP algorithm computes local conformity scores, sketches them with characterization vectors, and detects malicious clients based on averaged vector distance. Clients deemed highly malicious are subsequently excluded from the calibration process. Furthermore, we provide a technique for estimating the number of malicious clients, when their exact count is unknown, by optimizing the likelihood of the characterization vectors. Our theoretical analysis of the coverage bounds shows that the coverage of Rob-FCP is asymptotically close to the desired coverage level as long as the number of malicious clients is less than that of benign clients and the sample sizes of benign clients are sufficiently large.

We empirically evaluate Rob-FCP against multiple Byzantine attacks. Rob-FCP outperforms FCP by a large margin and even achieves comparable prediction coverage and efficiency as the benign settings on *five* realistic datasets covering multiple fields. We also demonstrate the validity and tightness of the bounds of prediction coverage with different ratios of malicious clients. We further conduct a set of ablation studies on the methods of conformity scores characterization and different distance measurements to highlight the critical components in Rob-FCP.

Technical Contributions: Our contributions span both theoretical and empirical aspects.

- We provide the *first* certifiably robust federated conformal prediction framework (Rob-FCP) in the Byzantine setting where malicious clients can report arbitrary conformity score statistics.
- We propose a maliciousness score to effectively detect Byzantine clients and a malicious client number estimator to predict the number of Byzantine clients.
- We theoretically certify the coverage guarantees of Rob-FCP. We also theoretically analyze the precision of the malicious client number estimator.
- We empirically demonstrate the robustness of Rob-FCP in federated Byzantine settings across multiple datasets. We also empirically validate the soundness and tightness of the coverage guarantees.

2. Preliminaries

2.1. Conformal prediction

Suppose that we have n data samples $\{(X_i, Y_i)\}_{i=1}^n$ with features $X_i \in \mathbb{R}^d$ and labels $Y_i \in \mathcal{Y} := \{1, 2, ..., C\}$. Assume that the data samples are drawn exchangeably from some unknown joint distribution of feature X and label Y, denoted by $\mathcal{P}_{XY}$. Given a desired coverage $1-\alpha \in (0,1)$, conformal prediction methods construct a prediction set $\hat{C}_{n,\alpha} \subseteq \mathcal{Y}$ for a new data sample $(X_{n+1}, Y_{n+1}) \sim \mathcal{P}_{XY}$ with the guarantee of *marginal prediction coverage*: $\mathbb{P}[Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1})] \geq 1-\alpha$.

In this work, we focus on the split conformal prediction setting (Papadopoulos et al., 2002), where the data samples are randomly partitioned into two disjoint sets: a training set $\mathcal{I}_{\text{tr}}$ and a calibration (hold-out) set $\mathcal{I}_{\text{cal}} = [n] \backslash \mathcal{I}_{\text{tr}}$. [1] We fit a classifier to the training set $\mathcal{I}_{\text{tr}}$ to estimate the conditional class probability $\pi : \mathbb{R}^d \mapsto \Delta^C$, with the y-th element denoted as $\pi_y(x) = \mathbb{P}[Y = y | X = x]$. Using the estimated probabilities that we denote by $\hat{\pi}(x)$, we then compute a non-conformity score $S_{\hat{\pi}}(X_i, Y_i)$ for each sample in the calibration set $\mathcal{I}_{\text{cal}}$. The non-conformity score measures how much non-conformity each sample has with respect to its ground truth label. A small non-conformity score $S_{\hat{\pi}}(X_i, Y_i)$ indicates that the estimated class probability $\hat{\pi}(X_i)$ aligns well with the ground truth label Y_i for the data sample (X_i, Y_i). A simple and standard non-conformity score (Sadinle et al., 2019) is $S_{\hat{\pi}}(x, y) = 1 - \hat{\pi}_y(x)$.

Given a desired coverage $1-\alpha$, the prediction set of the new test data point X_{n+1} is formulated as:

$$\hat{C}_{n,\alpha}(X_{n+1}) = \{y \in \mathcal{Y} : S_{\hat{\pi}}(X_{n+1}, y) \leq Q_{1-\alpha}\left(\{S_{\hat{\pi}}(X_i, Y_i)\}_{i \in \mathcal{I}_{\text{cal}}}\right)\}, \quad (1)$$

where $Q_{1-\alpha}(\{S_{\hat{\pi}}(X_i, Y_i)\}_{i \in \mathcal{I}_{\text{cal}}})$ is the $\lceil (1-\alpha)(1 + |\mathcal{I}_{\text{cal}}|) \rceil$-th largest value of the set $\{S_{\hat{\pi}}(X_i, Y_i)\}_{i \in \mathcal{I}_{\text{cal}}}$. The prediction set $\hat{C}_{n,\alpha}(X_{n+1})$ includes all the labels with a smaller non-conformity score than the $(1-\alpha)$-quantile of scores in the calibration set. Since we assume the data samples are exchangeable, the marginal coverage of the prediction set $\hat{C}_{n,\alpha}(X_{n+1})$ is no less than $1-\alpha$. We refer to (Vovk et al., 2005) for a more rigorous analysis of the prediction coverage.

2.2. Federated conformal prediction

In federated learning, multiple clients own their private data locally and collaboratively develop a global model.

[1] In here and what follows, $[n] := \{1, \cdots, n\}$.

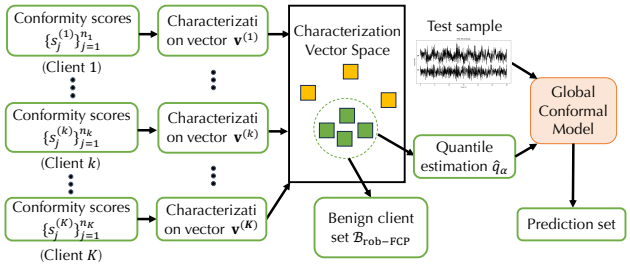


Figure 2: Overview of Rob-FCP.

Let K be the number of clients. We denote the local data distribution of the k-th client ($k \in [K]$) by $\mathcal{P}^{(k)}$. Let $\{(X_i^{(k)}, Y_i^{(k)})\}_{i\in[n_k]} \sim \mathcal{P}^{(k)}$ be n_k calibration samples owned by the k-th client. We denote $(X_{\text{test}}, Y_{\text{test}})$ as the future test point sampled from the global distribution $\mathcal{Q}_{\text{test},\lambda}$ for some probability vector $\lambda \in \Delta^K$: $(X_{\text{test}}, Y_{\text{test}}) \sim \mathcal{Q}_{\text{test},\lambda} := \sum_{k=1}^K \lambda_k \mathcal{P}^{(k)}$. Let $N = \sum_{k=1}^K n_k$ be the total sample size of K clients and $\hat{q}_\alpha$ be the $\lceil (1-\alpha)(N+K) \rceil$-th largest value in $\{S_{\hat{\pi}}(X_i^{(k)}, Y_i^{(k)})\}_{i\in[n_k],k\in[K]}$, where $\hat{\pi}$ is the collaboratively trained conditional class probability estimator ($\alpha \geq 1/(N/K+1)$). FCP (Lu et al., 2023) proves that under the assumption of partial exchangeability (Carnap & Jeffrey, 1980) and $\lambda_k \propto (n_k+1)$, the prediction set $\hat{C}_\alpha(X_{\text{test}}) = \{y \in \mathcal{Y} : S_{\hat{\pi}}(X_{\text{test}}, y) \leq \hat{q}_\alpha\}$ is a valid conformal prediction set with the guarantee:

$$1-\alpha \leq \mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] \leq 1-\alpha+\frac{K}{N+K}. \quad (2)$$

Considering communication cost and privacy concerns, having all agents upload their local non-conformity scores to the server for quantile computation of $\hat{q}_\alpha$ is impractical. Consequently, FCP (Lu et al., 2023) utilizes data sketching algorithms like T-digest (Dunning, 2021) for efficient and privacy-preserving distributed quantile estimation. They prove that if the rank of quantile estimate $\hat{q}_\alpha$ is between $(1-\alpha-\epsilon)(N+K)$ and $(1-\alpha+\epsilon)(N+K)$ where ϵ *denotes the quantile estimation error induced by data sketching*, then the guarantee in Equation (2) can be corrected as the following:

$$1-\alpha-\frac{\epsilon N+1}{N+K} \leq \mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] \leq 1-\alpha+\epsilon+\frac{K}{N+K}, \quad (3)$$

where K is the number of clients and N is the total sample sizes of clients.

3. Rob-FCP and coverage guarantees

3.1. Threat model

We follow the standard setup of FCP in Section 2.2 and consider the following Byzantine threat model. Suppose that among K clients, there exist K_b benign clients and K_m ($K_m = K - K_b$) malicious (Byzantine) clients. Without loss of generality, let the clients indexed by $[K_b] = \{1, ..., K_b\}$ be benign clients and the clients indexed by $[K]\backslash[K_b] = \{K_b+1, ..., K\}$ be malicious clients. The k-th benign client ($k \in [K_b]$) leverage the collaboratively trained global model $\hat{\pi}$ to compute the conformity scores on its local calibration data and sketched the score statistics with a characterization vector $\mathbf{v}^{(k)} \in \Delta^H$ where H is the granularity of the characterization statistics and then report the score vector $\mathbf{v}^{(k)}$ to the server. In contrast, K_m malicious clients can submit arbitrary characterization vectors $\mathbf{v}^{(k)} (k \in [K]\backslash[K_b])$ to the server.

Following FCP (Lu et al., 2023), the server considers a global distribution $\mathcal{Q}$ as a weighted combination of local distributions, denoted by $\mathcal{Q} = \sum_{i=1}^K \lambda_i \mathcal{P}^{(i)}$, where λ_i represents the weight assigned to each local distribution and is proportional to the size of local samples n_i: $\lambda_i \propto (n_i+1)$. Note that the server knows the true weights of local distributions (or equivalently, quantities of local samples), which can not be manipulated by malicious clients during the conformal prediction phase. Since the weights of local distributions (or equivalently, quantities of local samples) are a known priori to the server during the federated model learning phase, the threat model is reasonable and practical, aligning with the existing Byzantine analysis literature (Blanchard et al., 2017; Park et al., 2021; Data & Diggavi, 2021). For the threat model, we aim to develop a Byzantine-robust FCP framework (Rob-FCP) that maintains coverage and prediction efficiency despite the existence of malicious clients. We also aim to provide rigorous coverage guarantees of Rob-FCP in the Byzantine setting

3.2. Rob-FCP algorithm

Rob-FCP first detects the set of malicious clients, then excludes their score statistics during the computation of empirical quantile of conformity scores, and finally performs federated conformal prediction with the quantile value, which is not affected by malicious clients.

Characterization of conformity scores Let $\{s_j^{(k)}\}_{j\in[n_k]}$ be the conformity scores computed by the k-th client ($k \in [K]$) on its local calibration set. Since it is challenging to detect abnormal behavior from the unstructured and unnormalized conformity scores, we characterize the local conformity scores $\{s_j^{(k)}\}_{j\in[n_k]}$ with a vector $\mathbf{v}^{(k)} \in \mathbb{R}^H$ for client k, where the vector dimension $H \in \mathbb{Z}^+$ implicates the granularity of the characterization. Specifically, we can partition the range of conformity score values (e.g., $[0,1]$ for APS score (Romano et al., 2020)) into H subintervals $\{[a_h, a_{h+1})\}_{0\leq h\leq H-2} \cup \{[a_{H-1}, a_H]\}$, where a_h denotes the h-th cut point.[2] Thus, the h-th element of

[2]For simplicity, we abuse the last interval $[a_{H-1}, a_H]$ as $[a_{H-1}, a_H)$ in the future discussions.

the characterization vector ($\mathbf{v}_h^{(k)}$) represents the probability that a conformity score falls within the specific subinterval $[a_{h-1}, a_h)$:

$$\begin{aligned}\mathbf{v}_h^{(k)} &= \mathbb{P}_{s\sim\left\{s_j^{(k)}\right\}_{j\in[n_k]}}\left[a_{h-1} \leq s < a_h\right] \\ &= \frac{1}{n_k}\sum_{j=1}^{n_k}\mathbb{I}\left[a_{h-1} \leq s_j^{(k)} < a_h\right],\end{aligned} \tag{4}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. The characterization vector $\mathbf{v}^{(k)}$ is designed to encapsulate the distribution of score samples via histogram statistics, reflecting a fundamental multinomial distribution. This methodology leverages the observation that conformity scores originating from homogeneous distributions typically show substantial similarity. Consequently, characterization vectors from benign clients exhibit notable resemblance, in contrast to those from malicious clients, whose score statistics are anomalous. Such a distinct pattern facilitates the reliable identification of malicious clients.

Furthermore, Rob-FCP is designed with the flexibility to incorporate various methodologies for representing empirical conformity score samples as a real-valued vector $\mathbf{v}$. Among these methodologies are kernel density estimation (Terrell & Scott, 1992), offering a more nuanced interpretation of histogram statistics; parametric model fitting, such as Gaussian models; and clustering-based exemplar representations, including KMeans. The empirical analysis in Section 5.3 indicates that the histogram-based approach, as formulated in Equation (4), surpasses both parametric models and clustering techniques in performance. Hence, we consider the histogram statistic in Rob-FCP as our primary method of analysis.

Maliciousness score computation Rob-FCP detects the malicious clients via a maliciousness score in the space of characterization vectors. First, we compute pairwise ℓ_p ($p \in \mathbb{Z}^+$) vector distances among K clients:

$$d_{k_1,k_2} = \|\mathbf{v}^{(k_1)} - \mathbf{v}^{(k_2)}\|_p,\ \forall k_1, k_2 \in [K]. \tag{5}$$

Denote $N_{ear}(k,t)$ as the index set of the t-nearest neighbors of client k (excluding itself), with the distance between two clients k_1 and k_2 given by Equation (5). We define the maliciousness score $M(k) \in \mathbb{R}$ of client k ($k \in [K]$) as the averaged distance to the $K_b - 1$ nearest neighbors, where K_b is the number of benign clients:

$$M(k) = \frac{1}{K_b - 1}\sum_{k' \in N_{ear}(k, K_b-1)} d_{k,k'}. \tag{6}$$

We define the benign set identified by Rob-FCP, denoted as $\mathcal{B}_{\text{Rob-FCP}}$, as the set containing the indices of clients with the lowest K_b maliciousness scores among $\{M(k)\}_{k=1}^K$. Subsequently, quantile estimation $\hat{q}_\alpha$ is carried out using the characterization vectors from the clients within the benign set $\mathcal{B}_{\text{Rob-FCP}}$. The quantile estimation $\hat{q}\alpha$ is then applied to perform federated conformal prediction on the globally trained model in a distributed manner. An overview of Rob-FCP is presented in Figure 2, with the pseudocode detailed in Algorithm 1 in Appendix F.

To impair the overall performance of global conformal predictions, malicious clients often submit conformity score statistics that starkly contrast with those of benign clients. This difference results in the characterization vectors of malicious clients being distinct and separable from the aggregation of benign vectors. The calculation of maliciousness scores, which is based on the average distance to the $K_b - 1$ nearest neighbors, further accentuates this separation. Specifically, malicious clients tend to have higher maliciousness scores than benign clients, given the condition $K_b > K_m$, a common assumption in Byzantine resilience studies (Blanchard et al., 2017)). Leveraging this distinction, Rob-FCP effectively isolates and disregards the skewed statistics introduced by malicious clients during the conformal calibration process, thereby maintaining the validity of the conformal prediction set. A theoretical analysis of Rob-FCP, including rigorous coverage bounds, is provided in Section 3.3.

Effectiveness of Rob-FCP against mimick attacks Malicious clients with mimic attack (Karimireddy et al., 2022; Shejwalkar & Houmansadr, 2021) transmit similar gradients to benign clients in FL optimization, which is stealthy and deteriorates the optimization process by overrepresenting the mimicked clients in the setting with high data heterogeneity. However, in FCP, for a collaboratively trained model, we observe that the heterogeneity of distributions of non-conformity scores cannot be effectively used by mimic attacks to disturb the FCP process. The major difference between the setting in (Karimireddy et al., 2022; Shejwalkar & Houmansadr, 2021) and FCP is that the former considers the FL optimization, where clients perform multi-step local updates on local data distribution, and thus the gradients among clients can show a pretty high heterogeneity due to the data heterogeneity and also the high dimensionality of the gradients. This makes a great opportunity for the attackers to hide in and still distort the FL optimization effectively. However, in the FCP setting, the model is well-trained and converges well. Thus, the heterogeneity in the space of nonconformity score vectors is not as great as the heterogeneity in the high-dimensional gradient space during optimization. Note that in Rob-FCP, we do not have assumptions that the malicious clients should be very different from benign clients. The principle of the effectiveness of Rob-FCP is that (1) if the score vector of malicious clients is close to the benign clients, although Rob-FCP may identify it as benign, it can only make a limited and bounded difference on the FCP results,

and (2) if the score vector is far from the benign cluster, although it is effective to distort FCP, Rob-FCP will filter it out in this case. We provide the empirical validation results of the observation in Table 7 in Appendix G.2.

3.3. Coverage guarantee of Rob-FCP

We rigorously analyze the lower and upper bounds of the prediction coverage of Rob-FCP in the Byzantine setting in Theorem 1. The analysis reveals that, with an adequately large sample size of benign clients, Rob-FCP is capable of reaching the desired coverage level. This finding underscores the effectiveness of Rob-FCP in maintaining reliable prediction coverage, even in the presence of Byzantine clients.

Theorem 1 (Coverage guarantees of Rob-FCP in Byzantine setting)**.** *Consider FCP setting with K_b benign clients and K_m malicious clients. The k-th client reports the characterization vector $\mathbf{v}^{(k)}$ and local sample size n_k to the server ($k \in [K_b + K_m]$). Assume that the benign characterization vector $\mathbf{v}^{(k)}$ follows multinomial distribution $\mathcal{D}_k$ with event probability $\overline{\mathbf{v}}^{(k)}$ for the k-th client ($k \in [K_b]$). We use σ to quantify the heterogeneity of benign vectors as $\sigma = \max_{k_1 \in [K_b], k_2 \in [K_b]} \|\overline{\mathbf{v}}^{(k_1)} - \overline{\mathbf{v}}^{(k_2)}\|_1$. Let ϵ be the data sketching error as Equation* (3)*. Under the assumption that $K_m < K_b$, the following coverage guarantee for test instance (X_t, Y_t) holds with probability $1 - \beta$:*

$$\mathbb{P}\Big[Y_t \in \hat{C}_\alpha(X_t)\Big] \geq 1 - \alpha - P_{\text{byz}} - \frac{N_m \sigma}{n_b(1-\tau)} - \frac{\epsilon n_b + 1}{n_b + K_b}$$

$$\mathbb{P}\Big[Y_t \in \hat{C}_\alpha(X_t)\Big] \leq 1 - \alpha + P_{\text{byz}} + \frac{N_m \sigma}{n_b(1-\tau)} + \frac{\epsilon n_b + (\epsilon+1)K_b}{n_b + K_b}$$

$$\textit{where} \quad P_{\text{byz}} = \frac{H\Phi^{-1}(1 - \beta/2HK_b)}{2\sqrt{n_b}} \left(1 + \frac{N_m}{n_b} \frac{2}{1-\tau}\right) \tag{7}$$

where $\tau = K_m/K_b$ is the ratio between the number of malicious clients and the number of benign clients, $N_m := \sum_{k \in [K] \setminus [K_b]} n_k$ is the total sample size of malicious clients, $n_b := \min_{k' \in [K_b]} n_{k'}$ is the minimal sample size of benign clients, and $\Phi^{-1}(\cdot)$ denotes the inverse of the cumulative distribution function (CDF) of standard normal distribution.

Remark. **(R1)** Equation (7) offers the lower and upper bound of the prediction coverage with Rob-FCP in the Byzantine setting. The coverage bounds are in relation to (a) Byzantine coverage penalty P_{byz}, (b) client disparity penalty $N_m\sigma/n_b(1-\tau)$, and (c) data sketching penalty $\epsilon n_b+1/n_b+K_b$ or $\epsilon n_b+(\epsilon+1)K_b/n_b+K_b$. **(R2)** The Byzantine coverage penalty P_{byz} is induced by the presence of malicious clients. It can be exacerbated by a large ratio of malicious clients (a large τ) and a large total sample size of malicious clients (a large N_m). However, the Byzantine coverage penalty P_{byz} can be effectively reduced by a larger benign sample size n_b. **(R3)** The client disparity penalty is induced by the data heterogeneity among clients. Similarly, it can be exacerbated by a large τ and N_m, but reduced by a large n_b. We leverage the maximal pairwise vector norm to quantify the client heterogeneity, which aligns with existing Byzantine analysis (Park et al., 2021; Data & Diggavi, 2021). **(R4)** The data sketching penalty is induced by the local approximation error ϵ as Equation (3), with more details provided in (Lu et al., 2023). **(R5)** The assumption $K_m < K_b$ ($i.e., \tau < 1$) requires that the number of malicious clients is less than the number of benign clients, aligning with the break point of $\lceil K/2 \rceil$ in Byzantine analysis (Blanchard et al., 2017; Yin et al., 2018; Guerraoui et al., 2018). **(R6)** There exists a trade-off of selecting the characterization granularity H. According to FCP (Lu et al., 2023), with the histogram estimate, when H decreases, the data sketching becomes rough and increases the approximation error ϵ. At the same time, a smaller H will decrease the Byzantine coverage penalty P_{byz} due to a better concentration rate. We empirically perform ablation studies on the selection of H in Appendix G.2. **(R7)** We bound the concentration of the characterization vectors with the binomial proportion confidence interval (Wallis, 2013). We also provide results with more advanced concentration bounds DKW inequality (Dvoretzky et al., 1956) in Appendix D. **(R8)** Asymptotically, as long as the benign sample size n_b is sufficiently large, both the coverage lower bound and the upper bound reach the desired coverage level $1-\alpha$, demonstrating the robustness of Rob-FCP.

Proof sketch. We first leverage statistical confidence intervals and union bounds to conduct concentration analysis of the characterization vectors $\mathbf{v}^{(k)}$ for benign clients ($1 \leq k \leq K_b$). Then we consider the maliciousness scores of critical clients and relax the histogram statistics error. We finally translate the error of aggregated statistics to the error of the coverage bounds by algebra analysis. We provide complete proofs in Appendix C.1.

4. Rob-FCP with unknown numbers of malicious clients

4.1. Malicious client number estimator

In the standard Byzantine framework (Blanchard et al., 2017; Park et al., 2021; Liu et al., 2023), the defender is often assumed to have prior knowledge of the quantity of malicious clients K_m. This number plays a pivotal role in defense strategies: underestimating it results in the inclusion of malicious clients, leading to a degradation in overall performance, while overestimating it results in the exclusion of benign clients, thereby causing a shift in the global data distribution. However, in real-world applications, the exact count of malicious clients is typically unknown to the server. To address this gap and enhance the system's resilience in more complex Byzantine environments where the number of malicious clients is uncertain, we introduce a

novel estimator for malicious client numbers for Rob-FCP.

To accurately estimate the number of malicious clients K_m, we pivot to calculating the number of benign clients K_b, given the total client count K is known. To achieve this, we aim to maximize the likelihood of benign characterization vectors while minimizing the likelihood of malicious characterization vectors over the number of benign clients $\hat{K}_b$. The likelihood computation necessitates a predefined distribution for benign characterization vectors.

Considering that benign characterization vectors $\mathbf{v}^{(k)}$ ($k \in [K_b]$) are sampled from a multinomial distribution, which, for substantial sample sizes, can be closely approximated by a multivariate normal distribution as (Severini, 2005), we proceed under the assumption that the benign characterization vectors are samples from a multivariate normal distribution denoted as $\mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^H$ represents the mean, and $\Sigma \in \mathbb{R}^{H\times H}$ denotes the covariance matrix.

Then, we use expectation–maximization (EM) algorithm to effectively estimate the number of benign clients $\hat{K}_b$. In the expectation (E) step, given the current estimate of benign client number $\tilde{K}_b$, we compute the expected Gaussian mean and covariance by the observations of benign characterization vectors, which can be identified by the Rob-FCP algorithm in Section 3.2. In the maximization (M) step, we maximize the likelihood of characterization vectors given the estimated Gaussian mean and covariance in the E step. Formally, let $I(\cdot) : [K] \mapsto [K]$ be the mapping from the rank of maliciousness scores by Rob-FCP to the client index. The EM optimization step can be formulated as:

$$\begin{aligned}\hat{K}_b = \arg\max_{z\in[K]} \Bigg[&\frac{1}{z}\sum_{k=1}^{z} \log p(\mathbf{v}^{(I(k))}; \hat{\mu}(z), \hat{\Sigma}(z)) \\ &- \frac{1}{K-z}\sum_{k=z+1}^{K} \log p(\mathbf{v}^{(I(k))}; \hat{\mu}(z), \hat{\Sigma}(z)) \Bigg]\end{aligned} \tag{8}$$

where $\hat{\mu}(z)$ and $\hat{\Sigma}(z)$ are the expected mean and covariance: $\hat{\mu}(z) = 1/z \sum_{k\in[z]} \mathbf{v}^{(I(k))}$, $\hat{\Sigma}(z) = \mathbb{E}_{k\in[z]}[(\mathbf{v}^{(I(k))} - \hat{\mu}(z))^T(\mathbf{v}^{(I(k))} - \hat{\mu}(z))]$, and $p(\mathbf{v}; \mu, \Sigma)$ computes the likelihood of $\mathbf{v}$ given Gaussian $\mathcal{N}(\mu, \Sigma)$ as $p(\mathbf{v}; \mu, \Sigma) = \exp\left(-1/2(\mathbf{v}-\mu)^T\Sigma^{-1}(\mathbf{v}-\mu)\right)/\sqrt{(2\pi)^H|\Sigma|}$. The EM optimization in Equation (8) essentially searches for $\hat{K}_b$ such that the characteristic vectors of $\hat{K}_b$ clients with the lowest maliciousness scores (higher probability of being benign) exhibit a strong alignment with the benign normal distribution, and conversely, the characteristic vectors of the remaining clients (more likely to be malicious) show a decreased likelihood of fitting the benign normal distribution. Note that the derived estimate of $\hat{K}_b$ can be utilized as the input parameter $\tilde{K}_b$ in subsequent iterations, allowing for the refinement of the estimation through recursive applications of the EM optimization process.

4.2. Precision of malicious client number estimator

In this part, we theoretically show the precision of benign client number estimate in Equation (8).

Theorem 2 (Precision of malicious client number estimator)**.** *Assume $\mathbf{v}^{(k)}$ ($k \in [K_b]$) are IID sampled from Gaussian $\mathcal{N}(\mu, \Sigma)$ with mean $\mu \in \mathbb{R}^H$ ($H \geq 2$) and positive definite covariance matrix $\Sigma \in \mathbb{R}^{H\times H}$. Let $d = \min_{k\in[K]\setminus[K_b]} \|\mathbf{v}^{(k)} - \mu\|_2$. Consider EM optimization as Equation* (8) *and an initial guess of benign client number $\tilde{K}_b$ such that $K_m < \tilde{K}_b \leq K_b$. Then we have:*

$$\begin{aligned}\mathbb{P}\left[\hat{K}_m = K_m\right] \geq &1 - \frac{(3\tilde{K}_b - K_m - 2)^2 Tr(\Sigma)}{(\tilde{K}_b - K_m)^2 d^2} \\ &- \frac{2(K+K_b)Tr(\Sigma)\sigma_{max}^2(\Sigma^{-1/2})}{\sigma_{min}^2(\Sigma^{-1/2})d^2}\end{aligned} \tag{9}$$

where $\sigma_{max}(\Sigma^{-1/2})$, $\sigma_{min}(\Sigma^{-1/2})$ denote the maximal and minimal eigenvalue of matrix $\Sigma^{-1/2}$, and $Tr(\Sigma)$ denotes the trace of matrix Σ.

Remark. **(R1)** The lower bound in Equation (9) rises as the minimal distance between the malicious characterization vector to the benign mean μ (i.e., d) increases. The lower bound asymptotically approaches 1 with a sufficiently large d. It implies that when the malicious characterization vector is far away from the benign cluster (i.e., a large d), the malicious client number estimator has a high precision. **(R2)** The lower bound in Equation (9) also shows that when the initial guess $\tilde{K}_b$ is closer to K_b, the lower bound of estimate precision is higher, demonstrating the effectiveness of iterative EM optimization with Equation (8). **(R3)** Note that the condition of the initial guess $K_m < \tilde{K}_b < K_b$ is satisfiable by simply setting $\tilde{K}_b = \lceil K/2 \rceil$.

Proof sketch. We first analyze the tail bound of the multivariate normal distribution as (Vershynin, 2018), and then derive the probabilistic relationships between the maliciousness scores of benign clients and those of malicious clients using the tail bounds. We finally upper bound the probability of overestimation and underestimation by opening up the probability formulations. We defer the complete proof to Appendix C.2.

5. Experiments

5.1. Experiment setup

Datasets We evaluate Rob-FCP on a variety of standard datasets, including MNIST (Deng, 2012), CIFAR-10 (Krizhevsky et al.), and Tiny-ImageNet (Le & Yang, 2015). Our evaluation of Rob-FCP also cover two realistic healthcare datasets: the Sleep Heart Health Study (SHHS) dataset (Zhang et al., 2018) and a pathology dataset PathMNIST (Yang et al., 2023).

Table 1: Marginal coverage / average set size under different Byzantine attacks with 40% ($K_m/K = 40\%$) malicious clients. The desired marginal coverage is 0.9. The Dirichlet parameter β is 0.5. Results that more closely align with those observed in an all-benign-client scenario (provided in Table 3 in Appendix G.2) are highlighted in bold.

Byzantine Attack Method	Coverage Attack		Efficiency Attack		Gaussian Attack	
	FCP	Rob-FCP	FCP	Rob-FCP	FCP	Rob-FCP
MNIST	0.805 / 1.284	**0.899 / 1.783**	1.000 / 10.00	**0.902 / 1.804**	0.941 / 2.227	**0.923 / 2.182**
CIFAR-10	0.829 / 1.758	**0.897 / 2.319**	1.000 / 10.00	**0.892 / 2.351**	0.970 / 3.863	**0.921 / 2.623**
Tiny-ImageNet	0.825 / 27.84	**0.903 / 43.47**	1.000 / 200.0	**0.904 / 43.68**	0.942 / 61.50	**0.928 / 54.91**
SHHS	0.835 / 1.095	**0.901 / 1.365**	1.000 / 6.000	**0.901 / 1.366**	0.937 / 1.609	**0.900 / 1.359**
PathMNIST	0.837 / 1.055	**0.900 / 1.355**	1.000 / 9.000	**0.900 / 1.344**	1.000 / 6.935	**0.926 / 1.585**

Data partition in federated conformal prediction Our approach of data partition adheres to the standard federated learning evaluation framework by using the Dirichlet distribution to create different label ratios across clients (Yurochkin et al., 2019; Lin et al., 2020; Wang et al., 2020; Gao et al., 2022). Concretely, we sample $p_{c,j} \sim \text{Dir}(\beta)$ and allocate a portion of $p_{c,j}$ instances with class c to the client j, where $\text{Dir}(\cdot)$ denotes the Dirichlet distribution and β is a concentration parameter ($\beta > 0$), controlling the degree of data heterogeneity among clients. A lower β value results in a more heterogeneous data distribution. By default, we set β to 0.5 to establish a consistent level of data heterogeneity. Additionally, we explore alternative methods for generating heterogeneous data that reflect demographic variations. We segment the SHHS dataset based on five attributes (wake time, N1, N2, N3, REM), distributing instances to clients based on differing attribute intervals, thereby introducing another dimension of data diversity.

Byzantine attacks To evaluate the robustness of Rob-FCP in the Byzantine setting, we conducted comparisons with the baseline FCP (Lu et al., 2023) under three types of Byzantine attacks: (1) *coverage attack* (CovAttack) involves malicious clients reporting maximized conformity scores (e.g., 1 for LAC score (Sadinle et al., 2019)) to artificially inflate the conformity score at the targeted quantile, resulting in reduced coverage; (2) *efficiency attack* (EffAttack) involves malicious clients submitting minimized conformity scores (e.g., score of 0 for LAC score) to lower the conformity score at the quantile, thereby expanding the prediction set; (3) Gaussian Attack (GauAttack) involves malicious clients dispersing random Gaussian noise with a standard deviation of 0.5 into the scores, thereby disrupting the conformal calibration process.

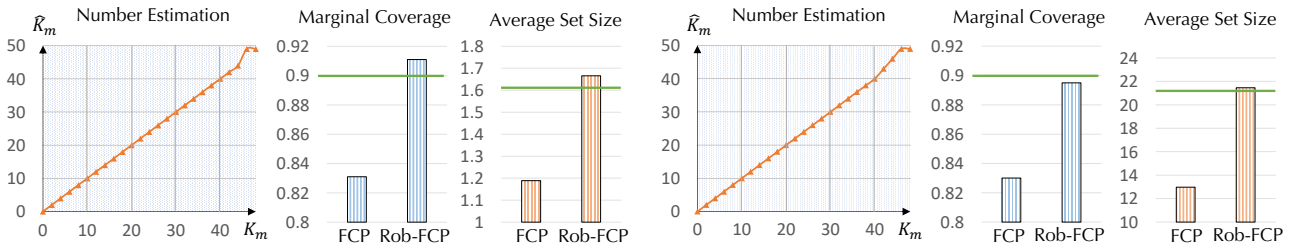
Evaluation metric We consider the global test data set $\mathcal{D}_{\text{test}} = \{(X_i, Y_i)\}_{i=1}^{N_{\text{test}}}$. We notate $C_\alpha(X_i)$ as the conformal prediction set given test sample X_i and consider the desired coverage level $1-\alpha$. We evaluate with the metrics of *marginal coverage* $\sum_{i=1}^{N_{\text{test}}} \mathbb{I}\left[Y_i \in C_\alpha(X_i)\right] / N_{\text{test}}$ and *average set size* $\sum_{i=1}^{N_{\text{test}}} |C_\alpha(X_i)| / N_{\text{test}}$. Without specification, the desired coverage level $1-\alpha$ is set 0.9. We provide more details of experiment setups in Appendix G.1.

The codes to reproduce all the evaluation results are publicly available at `https://github.com/kangmintong/Rob-FCP`.

5.2. Evaluation results

Byzantine robustness of Rob-FCP We evaluate Rob-FCP in terms of marginal coverage and average set size under coverage attack, efficiency attack, and Gaussian attacks, and compare these results against the baseline FCP. We present the results of FCP and Rob-FCP in the existence of 40% ($K_m/K = 40\%$) malicious clients on MNIST, CIFAR-10, Tiny-ImageNet, SHHS, and PathMNIST in Table 1. Under Byzantine attacks, FCP shows a significant deviation from the targeted coverage level of 0.9 and the expected benign set size. In contrast, Rob-FCP maintains comparable levels of marginal coverage and average set size, underlining its robustness. Note that while a smaller prediction set is generally preferred for efficiency, the primary objective here is to accurately meet the targeted coverage level of 0.9. Further results on the resilience of Rob-FCP against different portions of malicious clients (30%, 20%, and 10%) are provided in Table 9 in Appendix G.2. In Table 7 of Appendix G.2, we empirically demonstrate the robustness of Rob-FCP against mimick attacks (Karimireddy et al., 2022), which operate under a more restricted threat model that relies on knowing the score statistics of agents.

Rob-FCP with unknown numbers of malicious clients In Section 4, we explore a complex Byzantine scenario where the exact count of malicious clients is not known to the defender. To address this challenge, we introduce an estimator designed to predict the number of malicious participants accurately, with theoretical guarantee as Theorem 2. Our evaluation focuses on assessing the precision of this malicious client number estimator and examining the conformal prediction performance of Rob-FCP within this uncertain environment. The results in Figure 3 reveal that our estimator ($\hat{K}_m$) closely approximates the actual number of malicious clients (K_m), leading to a marginal coverage and average set size close to the benign level. Further results across all five datasets, under a variety of Byzantine



(a) CIFAR-10 (b) Tiny-ImageNet

Figure 3: Results of malicious client number estimation and conformal prediction performance in the setting with unknown numbers of malicious clients. The green horizontal line denotes the benign conformal performance. Rob-FCP estimates the number of malicious clients faithfully, and provides an empirical coverage rate matching the target (benign level).

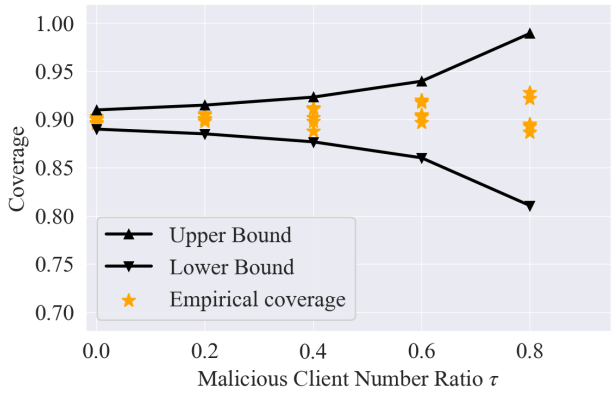


Figure 4: Upper and lower bounds of prediction coverage of Rob-FCP by Theorem 1 on Tiny-ImageNet.

attacks, are detailed in Table 10 in Appendix G.2, confirming the effectiveness of the malicious client number estimator.

Validation of coverage bounds of Rob-FCP In Theorem 1, we provide both the lower and upper bound of the coverage rate of Rob-FCP, considering the ratio of malicious clients ($\tau = K_m/K_b$) and the sample sizes across clients. In Figure 4, we compare these theoretical bounds of coverage rate against the observed empirical marginal coverage under Gaussian attacks on Tiny-ImageNet. The results demonstrate the validity and tightness of the certified coverage bounds in Theorem 1.

5.3. Ablation study

Robustness of Rob-FCP across varying levels of data heterogeneity Data heterogeneity among clients poses significant challenges to achieving precise federated conformal prediction. To assess the resilience of Rob-FCP to this issue, we conducted evaluations using various values of the Dirichlet parameter β, which modulates the degree of data heterogeneity among clients. The results in Table 2 show that Rob-FCP reliably maintains marginal coverage and average set size at levels close to the benign levels, underscoring its robustness in the face of data heterogeneity. Furthermore, we investigate additional approaches to create heterogeneous data that mirror demographic differences. This involves dividing the SHHS dataset according to five specific attributes (wake time, N1, N2, N3, REM) and allocating instances to clients based on varying intervals of these attributes. The results in Table 4 in Appendix G.2 highlight Rob-FCP's capability to effectively handle diverse forms of data heterogeneity.

Ablation study on conformity score distribution characterization methods A pivotal aspect of Rob-FCP involves the characterization of the conformity score distribution through empirical data. Our primary method utilizes histogram statistics as outlined in Equation (4). Alternatively, one could represent score samples using cluster centers derived from clustering algorithms like KMeans, or employ a parametric method such as fitting the score samples to a Gaussian distribution and characterizing them by the mean and variances of the Gaussian. Our empirical comparison of these methods, presented in Figure 5 and Figure 12 within Appendix G.2, reveals that the histogram statistics approach yields superior performance. Additional ablation studies focusing on various distance measurement techniques are provided in Figure 13 in Appendix G.2.

Robustness of Rob-FCP with various conformity scores Besides applying LAC nonconformity scores, we also evaluate Rob-FCP with APS conformity scores (Romano et al., 2020). The results in Figures 6 to 11 in Appendix G.2 demonstrate the resilience of Rob-FCP to the selection of conformity scores. We also evaluate the runtime and show the efficiency of Rob-FCP in Table 5 in Appendix G.2.

6. Related work

Conformal prediction is a statistical tool to construct the prediction set with guaranteed prediction coverage (Jin et al., 2023; Solari & Djordjilović, 2022; Yang & Kuchibhotla, 2021; Romano et al., 2020; Barber et al., 2021; Kang et al., 2024b;a), assuming exchangeable data. Recently, **federated conformal prediction** (FCP) (Lu & Kalpathy-Cramer, 2021; Lu et al., 2023) adapts the conformal predic-

Table 2: Marginal coverage / average set size across varying levels of data heterogeneity, controlled by different Dirichlet parameter β. The evaluation is done under coverage attack with 40% ($K_m/K = 40\%$) malicious clients. The desired coverage level is 0.9. Results that more closely align with those observed in an all-benign-client scenario (provided in Table 3) are highlighted in bold.

Dataset	Method	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 0.9$
MNIST	FCP	0.780 / 1.173	0.817 / 1.318	0.833 / 1.384	0.805 / 1.265	0.828 / 1.363
	Rob-FCP	**0.899 / 1.806**	**0.905 / 1.809**	**0.903 / 1.827**	**0.898 / 1.781**	**0.893 / 1.768**
CIFAR-10	FCP	0.806 / 1.641	0.821 / 1.717	0.836 / 1.791	0.823 / 1.744	0.824 / 1.723
	Rob-FCP	**0.899 / 2.260**	**0.907 / 2.405**	**0.892 / 2.243**	**0.904 / 2.396**	**0.910 / 2.416**
Tiny-ImageNet	FCP	0.840 / 28.625	0.830 / 28.192	0.833 / 28.340	0.821 / 27.140	0.831 / 28.751
	Rob-FCP	**0.913 / 45.872**	**0.910 / 44.972**	**0.898 / 42.571**	**0.887 / 41.219**	**0.898 / 43.298**
PathMNIST	FCP	0.850 / 1.106	0.839 / 1.065	0.837 / 1.055	0.839 / 1.065	0.832 / 1.043
	Rob-FCP	**0.895 / 1.311**	**0.900 / 1.355**	**0.900 / 1.355**	**0.899 / 1.354**	**0.901 / 1.363**

tion to the federated learning and provides a rigorous guarantee of the distributed uncertainty quantification framework. DP-FCP (Plassier et al., 2023) proposes federated CP with differential privacy guarantees and provides valid coverage bounds under label shifting among clients. Humbert et al. propose a quantile-of-quantiles estimator for federated conformal prediction with a one-round communication and provide a locally differentially private version. WFCP (Zhu et al., 2023) applies FCP to wireless communication. However, no prior works explore the robustness of FCP against Byzantine agents which can report malicious statistics to downgrade the conformal prediction performance. We are the first to propose a robust FCP method with valid and tight coverage guarantees.

Byzantine learning (Driscoll et al., 2003; Awerbuch et al., 2002; Lamport et al., 2019) refers to methods that can robustly aggregate updates from potentially malicious or faulty worker nodes in the distributed setting. Specifically, a line of works (Guerraoui et al., 2018; Pillutla et al., 2022; Data & Diggavi, 2021; Karimireddy et al., 2020; Yi et al., 2022) studies the resilience to Byzantine failures of distributed implementations of Stochastic Gradient Descent (SGD) and proposes different metrics to identify malicious gradients such as gradient norm (Blanchard et al., 2017) and coordinate-wise trimmed mean (Yin et al., 2018). However, the metrics are designed for the stability and convergence of distributed optimization and cannot be applied to the Byzantine FCP setting to provide rigorous coverage guarantees. In contrast, we propose Rob-FCP to perform Byzantine-robust distributed uncertainty quantification and provide valid and tight coverage bounds theoretically.

7. Conclusion

In this paper, we propose Rob-FCP, a certifiably Byzantine-robust federated conformal prediction algorithm with rigorous coverage guarantees. Rob-FCP sketches the local samples of conformity scores with characterization vectors and detects the malicious clients in the vector space. We empirically show the robustness of Rob-FCP against Byzantine failures on five datasets and validate the theoretical coverage bounds.

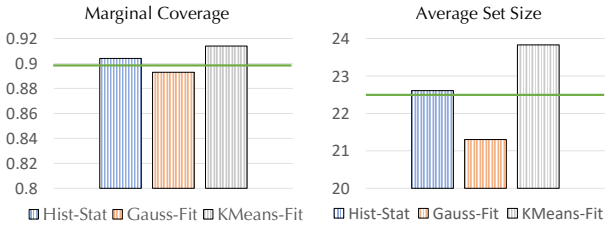


Figure 5: Marginal coverage / average set size under coverage attack with 40% malicious clients on Tiny-ImageNet. The green horizontal line denotes the benign marginal coverage and average set size without any malicious clients.

Acknowledgements

This work is supported by the National Science Foundation under grant No. 1910100, No. 2046726, No. 2229876, DARPA GARD, the National Aeronautics and Space Administration (NASA) under grant No. 80NSSC20M0229, the Alfred P. Sloan Fellowship, the Amazon research award, and the eBay research award. This work is supported by NSF award SCH-2205289, SCH-2014438, IIS-1838042, NIH award R01 1R01NS107291-01.

Impact Statement

We do not see potential ethical or societal issues about Rob-FCP. In contrast, Rob-FCP is a robust framework against malicious clients in the federated conformal prediction settings and can safeguard the applications of FCP in safety-critical scenarios such as healthcare and medical diagnosis.

References

Ahmad, M. A., Eckert, C., and Teredesai, A. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pp. 559–560, 2018.

Andrew, G., Thakkar, O., McMahan, B., and Ramaswamy, S. Differentially private learning with adaptive clipping. *Advances in Neural Information Processing Systems*, 34: 17455–17466, 2021.

Awerbuch, B., Holmer, D., Nita-Rotaru, C., and Rubens, H. An on-demand secure routing protocol resilient to byzantine failures. In *Proceedings of the 1st ACM workshop on Wireless security*, pp. 21–30, 2002.

Balasubramanian, V., Ho, S.-S., and Vovk, V. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. 2021.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., McMahan, B., et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1:374–388, 2019.

Carnap, R. and Jeffrey, R. C. *Studies in inductive logic and probability*, volume 2. Univ of California Press, 1980.

Data, D. and Diggavi, S. Byzantine-resilient high-dimensional sgd with local iterations on heterogeneous data. In *International Conference on Machine Learning*, pp. 2478–2488. PMLR, 2021.

Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Driscoll, K., Hall, B., Sivencrona, H., and Zumsteg, P. Byzantine fault tolerance, from theory to reality. In *International Conference on Computer Safety, Reliability, and Security*, pp. 235–248. Springer, 2003.

Dunning, T. The t-digest: Efficient estimates of distributions. *Software Impacts*, 7:100049, 2021. ISSN 2665-9638. doi: https://doi.org/10.1016/j.simpa.2020.100049. URL https://www.sciencedirect.com/science/article/pii/S2665963820300403.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

Erickson, B. J., Korfiatis, P., Akkus, Z., and Kline, T. L. Machine learning for medical imaging. *Radiographics*, 37(2):505–515, 2017.

Gao, L., Fu, H., Li, L., Chen, Y., Xu, M., and Xu, C.-Z. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10112–10121, 2022.

Guerraoui, R., Rouault, S., et al. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*, pp. 3521–3530. PMLR, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Humbert, P., Le Bars, B., Bellet, A., and Arlot, S. One-shot federated conformal prediction. In *International Conference on Machine Learning*, pp. 14153–14177. PMLR, 2023.

Jin, Y., Ren, Z., and Candès, E. J. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120, 2023.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Kang, M., Gürel, N. M., Li, L., and Li, B. Colep: Certifiably robust learning-reasoning conformal prediction via probabilistic circuits. *arXiv preprint arXiv:2403.11348*, 2024a.

Kang, M., Gürel, N. M., Yu, N., Song, D., and Li, B. C-rag: Certified generation risks for retrieval-augmented language models. *arXiv preprint arXiv:2402.03181*, 2024b.

Karimireddy, S. P., He, L., and Jaggi, M. Byzantine-robust learning on heterogeneous datasets via bucketing. *arXiv preprint arXiv:2006.09365*, 2020.

Karimireddy, S. P., He, L., and Jaggi, M. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=jXKKDEi5vJt.

Kompa, B., Snoek, J., and Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Lamport, L., Shostak, R., and Pease, M. The byzantine generals problem. In *Concurrency: the works of leslie lamport*, pp. 203–226. 2019.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Li, Q., Diao, Y., Chen, Q., and He, B. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 965–978. IEEE, 2022a.

Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022b.

Lin, T., Kong, L., Stich, S. U., and Jaggi, M. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33: 2351–2363, 2020.

Liu, Y., Chen, C., Lyu, L., Wu, F., Wu, S., and Chen, G. Byzantine-robust learning on heterogeneous data via gradient splitting. 2023.

Lu, C. and Kalpathy-Cramer, J. Distribution-free federated learning with conformal predictions. *arXiv preprint arXiv:2110.07661*, 2021.

Lu, C., Yu, Y., Karimireddy, S. P., Jordan, M. I., and Raskar, R. Federated conformal predictors for distributed uncertainty quantification. *arXiv preprint arXiv:2305.17564*, 2023.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

McMahan, H. B., Moore, E., Ramage, D., and y Arcas, B. A. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2:2, 2016.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pp. 345–356. Springer, 2002.

Park, J., Han, D.-J., Choi, M., and Moon, J. Sageflow: Robust federated learning against both stragglers and adversaries. *Advances in neural information processing systems*, 34:840–851, 2021.

Pillutla, K., Kakade, S. M., and Harchaoui, Z. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

Plassier, V., Makni, M., Rubashevskii, A., Moulines, E., and Panov, M. Conformal prediction for federated uncertainty quantification under label shift. *arXiv preprint arXiv:2306.05131*, 2023.

Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/244edd7e85dc81602b7615cd705545f5-Paper.pdf.

Sadinle, M., Lei, J., and Wasserman, L. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Severini, T. A. *Elements of distribution theory*, volume 17. Cambridge University Press, 2005.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Shejwalkar, V. and Houmansadr, A. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

Solari, A. and Djordjilović, V. Multi split conformal prediction. *Statistics & Probability Letters*, 184:109395, 2022.

Terrell, G. R. and Scott, D. W. Variable kernel density estimation. *The Annals of Statistics*, pp. 1236–1265, 1992.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wallis, S. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*, 20(3):178–208, 2013.

Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., and Khazaeni, Y. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10 (2):1–19, 2019.

Yang, Y. and Kuchibhotla, A. K. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 2021.

Yi, J., Wu, F., Zhang, H., Zhu, B., Qi, T., Sun, G., and Xie, X. Robust quantity-aware aggregation for federated learning. *arXiv preprint arXiv:2205.10848*, 2022.

Yin, D., Chen, Y., Kannan, R., and Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pp. 5650–5659. PMLR, 2018.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., and Khazaeni, Y. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pp. 7252–7261. PMLR, 2019.

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., and Redline, S. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

Zhang, X., Chen, X., Hong, M., Wu, Z. S., and Yi, J. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning, ICML 2022*, 2022.

Zheng, Q., Chen, S., Long, Q., and Su, W. Federated f-differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pp. 2251–2259. PMLR, 2021.

Zhu, M., Zecchin, M., Park, S., Guo, C., Feng, C., and Simeone, O. Federated inference with reliable uncertainty quantification over wireless channels via conformal prediction. *arXiv preprint arXiv:2308.04237*, 2023.

Contents

A Limitations and future works **14**

B Additional related work **14**

C Omitted proofs **14**

C.1 Proof of Theorem 1 14

C.2 Proof of Theorem 2 20

D Improvements with DKW inequality **23**

D.1 Improvement of Lemma C.1 with DKW inequality 23

D.2 Improvement of Theorem 1 with DKW inequality 24

E Analysis of Rob-FCP with an overestimated number of benign clients K'_b **24**

F Algorithm of Rob-FCP **26**

G Experiments **26**

G.1 Experiment setup 26

G.2 Additional evaluation results 27

A. Limitations and future works

One possible limitation of Rob-FCP may lie in the restriction of the targeted Byzantine threat model. We mainly consider the Byzantine setting where a certain ratio of malicious clients reports arbitrary conformity score statistics. In such a Byzantine case, the break point is $\lceil K/2 \rceil$, indicating that any algorithm cannot tolerate $\lceil K/2 \rceil$ or more malicious clients. However, in practice, malicious clients have the flexibility of only manipulating partial conformity scores. In this case, the potential break point is a function of the maximal ratio of manipulated scores for each client and can be larger than $\lceil K/2 \rceil$. Therefore, it is interesting for future work to analyze the break point of robust FCP algorithms with respect to the total manipulation sizes and budgets of manipulation sizes for each client. Another threat model worthy of exploration in future work is the adversarial setting in FCP. In the adversarial setting, malicious clients can only manipulate the data samples instead of the conformity scores to downgrade the FCP performance. Therefore, potential defenses can consider adversarial conformal training procedures to collaboratively train a robust FCP model against perturbations in the data space.

To provide differential privacy guarantees of Rob-FCP, one practical approach is to add privacy-preserving noises to the characterization vectors before uploading them to the server. Essentially, we can view the characterization vector as the gradient in the setting of FL with differential privacy (DP) and add Gaussian noises to the characterization vector with differential privacy guarantees as a function of the scale of noises, which can be achieved by drawing analogy from the FL with DP setting (Zheng et al., 2021; Andrew et al., 2021; Zhang et al., 2022). Therefore, practically implementing the differential-private version of Rob-FCP is possible and straightforward.

B. Additional related work

Byzantine learning (Driscoll et al., 2003; Awerbuch et al., 2002; Lamport et al., 2019) refers to methods that can robustly aggregate updates from potentially malicious or faulty worker nodes in the distributed setting. Specifically, a line of works (Guerraoui et al., 2018; Pillutla et al., 2022; Data & Diggavi, 2021; Karimireddy et al., 2020; Yi et al., 2022) studies the resilience to Byzantine failures of distributed implementations of Stochastic Gradient Descent (SGD) and proposes different metrics to identify malicious gradients such as gradient norm (Blanchard et al., 2017) and coordinate-wise trimmed mean (Yin et al., 2018). However, the metrics are designed for the stability and convergence of distributed optimization and cannot be applied to the Byzantine FCP setting to provide rigorous coverage guarantees. In contrast, we propose Rob-FCP to perform Byzantine-robust distributed uncertainty quantification and provide valid and tight coverage bounds theoretically.

C. Omitted proofs

C.1. Proof of Theorem 1

Before proving Theorem 1, we first prove the following lemma.

Lemma C.1. *For K clients including K_b benign clients and $K_m := K - K_b$ malicious clients, each client reports a characterization vector $\mathbf{v}^{(k)} \in \Delta^H$ $(k \in [K])$ and a quantity $n_k \in \mathbb{Z}^+$ $(k \in [K])$ to the server. Suppose that the reported characterization vectors of benign clients are sampled from the same underlying multinomial distribution $\mathcal{D}$, while those of malicious clients can be arbitrary. Let ϵ be the estimation error of the data sketching by characterization vectors as illustrated in Equation* (3)*. Under the assumption that $K_m < K_b$, the following holds with probability $1-\beta$:*

$$\begin{aligned}
\mathbb{P}\left[Y_{test} \in \hat{C}_\alpha(X_{test})\right] &\geq 1 - \alpha - \frac{\epsilon n_b + 1}{n_b + K_b} - \frac{H\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{n_b}}\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right), \\
\mathbb{P}\left[Y_{test} \in \hat{C}_\alpha(X_{test})\right] &\leq 1 - \alpha + \epsilon + \frac{K_b}{n_b + K_b} + \frac{H\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{n_b}}\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right).
\end{aligned} \tag{10}$$

where $\tau = K_m/K_b$ is the ratio of the number of malicious clients and the number of benign clients, $N_m := \sum_{k\in[K]\setminus[K_b]} n_k$ is the total sample size of malicious clients, $n_b := \min_{k'\in[K_b]} n_{k'}$ is the minimal sample size of benign clients, and $\Phi^{-1}(\cdot)$ denotes the inverse of the cumulative distribution function (CDF) of standard normal distribution.

Proof. The proof consists of 3 parts: (a) concentration analysis of the characterization vectors $\mathbf{v}^{(k)}$ for benign clients $(1 \leq k \leq K_b)$, (b) analysis of the algorithm of the identification of malicious clients, and (c) analysis of the error of the coverage bound.

Part (a): concentration analysis of the characterization vectors $\mathbf{v}^{(k)}$ for benign clients ($1 \leq k \leq K_b$).

Let $\mathbf{v}_h^{(k)}$ be the h-th element of vector $\mathbf{v}^{(k)}$. By definition, since $\mathbf{v}^{(k)}$ is sampled from a multinomial distribution, $\mathbf{v}_h^{(k)}$ denotes the success rate estimate of a Bernoulli distribution. We denote the event probabilities of the multinomial distribution $\mathcal{D}$ as $\overline{\mathbf{v}}$. Therefore, the true success rate of the Bernoulli distribution at the h-th position is $\overline{\mathbf{v}}_h$. According to the binomial proportion confidence interval (Wallis, 2013), we have:

$$\mathbb{P}\left[\left|\mathbf{v}_h^{(k)} - \overline{\mathbf{v}}_h\right| > \Phi^{-1}(1-\beta/2HK_b)\frac{\sqrt{n_{ks}n_{kf}}}{n_k\sqrt{n_k}}\right] \leq \beta/HK_b, \tag{11}$$

where β/HK_b is the probability confidence, $\Phi^{-1}(\cdot)$ denotes the inverse of the CDF of the standard normal distribution, and n_{ks} and $n_{kf} := n_k - n_{ks}$ are the number of success and failures in n_k Bernoulli trials, respectively. Applying the inequality $n_{ks}n_{kf} \leq n_k^2/4$ in Equation (11), the following holds:

$$\mathbb{P}\left[\left|\mathbf{v}_h^{(k)} - \overline{\mathbf{v}}_h\right| > \frac{\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{n_k}}\right] \leq \beta/HK_b. \tag{12}$$

Applying the union bound for H elements in vector $\mathbf{v}^{(k)}$ and K_b characterization vectors of benign clients, the following holds with probability $1-\beta$:

$$\left|\mathbf{v}_h^{(k)} - \overline{\mathbf{v}}_h\right| \leq \frac{\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{\min_{k'\in[K_b]} n_{k'}}},\ \forall k \in [K_b],\ \forall h \in [H], \tag{13}$$

from which we can derive the bound of difference for ℓ_1 norm distance as:

$$\left\|\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_1 \leq r(\beta) := \frac{H\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{\min_{k'\in[K_b]} n_{k'}}},\ \forall k \in [K_b], \tag{14}$$

where $r(\beta)$ is the perturbation radius of random vector $\mathbf{v}$ given confidence level $1-\beta$. $\forall k_1, k_2 \in [K_b]$, the following holds with probability $1-\beta$ due to the triangular inequality:

$$\left\|\mathbf{v}^{(k_1)} - \mathbf{v}^{(k_2)}\right\|_1 \leq \left\|\mathbf{v}^{(k_1)} - \overline{\mathbf{v}}\right\|_1 + \left\|\mathbf{v}^{(k_2)} - \overline{\mathbf{v}}\right\|_1 \leq 2r(\beta). \tag{15}$$

Furthermore, due to the fact that $\|\mathbf{v}\|_p \leq \|\mathbf{v}\|_1$ for any integer $p \geq 1$, the following holds with probability $1-\beta$:

$$\left\|\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_p \leq \left\|\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_1 \leq r(\beta), \tag{16}$$

$$\left\|\mathbf{v}^{(k_1)} - \mathbf{v}^{(k_2)}\right\|_p \leq \left\|\mathbf{v}^{(k_1)} - \mathbf{v}^{(k_2)}\right\|_1 \leq 2r(\beta). \tag{17}$$

Part (b): analysis of the algorithm of the identification of malicious clients.

Let $N(k,n)$ be the set of the index of n nearest clients to the k-th client based on the metrics of ℓ_p norm distance in the space of characterization vectors. Then the maliciousness scores $M(k)$ for the k-th client ($k \in [K]$) can be defined as:

$$M(k) := \frac{1}{K_b - 1}\sum_{k'\in N(k,K_b-1)}\left\|\mathbf{v}^{(k)} - \mathbf{v}^{(k')}\right\|_p. \tag{18}$$

Let $\mathcal{B}$ be the set of the index of benign clients identified by Algorithm 1 by selecting the clients associated with the lowest K_b maliciousness scores. We will consider the following cases separately: (1) $\mathcal{B}$ contains exactly K_b benign clients, and (2) $\mathcal{B}$ contains at least one malicious client indexed by m.

Case (1): $\mathcal{B}$ ($|\mathcal{B}| = K_b$) contains exactly K_b benign clients. We can derive as follows:

$$\left\|\sum_{k=1}^{K_b}\frac{n_k}{N_b}\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_p \leq \sum_{k=1}^{K_b}\frac{n_k}{N_b}\left\|\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_p \qquad \text{[triangular inequality]} \tag{19}$$

$$\leq \sum_{k=1}^{K_b}\frac{n_k}{N_b}r(\beta) \qquad \text{[by Equation (16)]} \tag{20}$$

$$= r(\beta), \tag{21}$$

where $N_b := \sum_{k\in[K_b]} n_k$ is the total sample size of benign clients.

Case (2): $\mathcal{B}$ ($|\mathcal{B}| = K_b$) contains at least one malicious client indexed by m. Since we assume $K_m < K_b$, there are at most $K_b - 1$ malicious clients in $\mathcal{B}$. Therefore, there is at least 1 benign client in $[K]\backslash\mathcal{B}$ indexed by b. We can derive the lower bound of the maliciousness score for the m-th client $M(m)$ as:

$$M(m) = \frac{1}{K_b - 1} \sum_{k'\in N(m,K_b-1)} \left\|\mathbf{v}^{(m)} - \mathbf{v}^{(k')}\right\|_p \tag{22}$$

$$\geq \frac{1}{K_b - 1} \sum_{k'\in N(m,K_b-1),k'\in[K_b]} \left\|\mathbf{v}^{(m)} - \mathbf{v}^{(k')}\right\|_p . \tag{23}$$

Since there are at least $K_b - K_m$ benign clients in $\mathcal{B}$ (there are at most K_m malicious clients in $\mathcal{B}$), there exists one client indexed by b_b ($b_b \in \mathcal{B}$) such that:

$$\left\|\mathbf{v}^{(m)} - \mathbf{v}^{(b_b)}\right\|_p \leq \frac{(K_b - 1)M(m)}{K_b - K_m} \tag{24}$$

We can derive the upper bound of the maliciousness score for the b-th benign client $M(b)$ as:

$$M(b) = \frac{1}{K_b - 1} \sum_{k'\in N(b,K_b-1)} \left\|\mathbf{v}^{(b)} - \mathbf{v}^{(k')}\right\|_p \tag{25}$$

$$\leq 2r(\beta) \qquad \text{[by Equation (17)]} \tag{26}$$

Since the m-th client is included in $\mathcal{B}$ and identified as a benign client, while the b-th client is not in $\mathcal{B}$, the following holds according to the procedure in Algorithm 1:

$$M(b) \geq M(m), \tag{27}$$

from which we can derive the following by combining Equation (24) and Equation (26):

$$\left\|\mathbf{v}^{(m)} - \mathbf{v}^{(b_b)}\right\|_p \leq \frac{(K_b - 1)2r(\beta)}{K_b - K_m} \tag{28}$$

Then, we can derive the upper bound of $\left\|\mathbf{v}^{(m)} - \overline{\mathbf{v}}\right\|_p$, $\forall m \in \mathcal{B}$ and $K_b < m \leq K$ as follows:

$$\left\|\mathbf{v}^{(m)} - \overline{\mathbf{v}}\right\|_p \leq \left\|\mathbf{v}^{(m)} - \mathbf{v}^{(b_b)}\right\|_p + \left\|\mathbf{v}^{(b_b)} - \overline{\mathbf{v}}\right\|_p \tag{29}$$

$$\leq \frac{2(K_b - 1)r(\beta)}{K_b - K_m} + r(\beta) \tag{30}$$

Finally, we can derive as follows:

$$\left\|\sum_{k\in\mathcal{B}} \frac{n_k}{N_\mathcal{B}} \mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_p \leq \sum_{k\in\mathcal{B}} \frac{n_k}{N_\mathcal{B}} \left\|\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_p \tag{31}$$

$$\leq \sum_{k\in\mathcal{B},k\in[K_b]} \frac{n_k}{N_\mathcal{B}} \left\|\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_p + \sum_{k\in\mathcal{B},k\in[K]\backslash[K_b]} \frac{n_k}{N_\mathcal{B}} \left\|\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_p \tag{32}$$

$$\leq \sum_{k\in\mathcal{B},k\in[K_b]} \frac{n_k}{N_\mathcal{B}} r(\beta) + \sum_{k\in\mathcal{B},k\in[K]\backslash[K_b]} \frac{n_k}{N_\mathcal{B}} \left[\frac{2(K_b - 1)r(\beta)}{K_b - K_m} + r(\beta)\right] \tag{33}$$

$$\leq r(\beta) + \sum_{k\in\mathcal{B},k\in[K]\backslash[K_b]} \frac{n_k}{N_\mathcal{B}} \frac{2(K_b - 1)r(\beta)}{K_b - K_m} \tag{34}$$

$$\leq r(\beta)\left(1 + \frac{N_m}{\min_{k'\in[K_b]} n_{k'}} \frac{2}{1-\tau}\right), \tag{35}$$

where $N_m := \sum_{k\in[K]\setminus[K_b]} n_k$ is the total sample size of malicious clients, $N_{\mathcal{B}}$ is the total sample size of clients in $\mathcal{B}$, and $\tau := \dfrac{K_m}{K_b}$ is the ratio of the number of malicious clients to the number of benign clients.

Combining *case (1)* and *case (2)*, we can conclude that:

$$\left\|\sum_{k\in\mathcal{B}} \frac{n_k}{N_{\mathcal{B}}}\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_p \leq \max\left\{1, 1 + \frac{N_m}{\min_{k'\in[K_b]} n_{k'}} \frac{2}{1-\tau}\right\} r(\beta) \tag{36}$$

$$= \left(1 + \frac{N_m}{\min_{k'\in[K_b]} n_{k'}} \frac{2}{1-\tau}\right) r(\beta) \tag{37}$$

Part (c): analysis of the error of the coverage bound. In this part, we attempt to translate the error of aggregated vectors induced by malicious clients to the error of the bound of marginal coverage. Let $F_1(q, \mathbf{v}) := \sum_{j=1}^{H} \mathbb{I}\left[a_j < q\right] \mathbf{v}_j$, where $q \in [0,1]$ and a_j is the j-th partition point used to construct the characterization vector $\mathbf{v} \in \Delta^H$. Let $F_2(q, \mathbf{v}) := \sum_{j=1}^{H} \mathbb{I}\left[a_{j-1} < q\right] \mathbf{v}_j$. Then by definition, we know that $F_1(q_\alpha, \overline{\mathbf{v}}) \leq \mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] \leq F_2(q_\alpha, \overline{\mathbf{v}})$, where q_α is the true $(1-\alpha)$ quantile value of the non-conformity scores, $\overline{\mathbf{v}}$ is the event probability of the multinormial distribution $\mathcal{D}$, and $\hat{C}_\alpha(X_{\text{test}})$ is the conformal prediction set of input X_{test} using the true benign calibrated conformity score q_α and statistics of score distribution $\overline{\mathbf{v}}$.

Let $\hat{q}_\alpha$ be the quantile estimate during calibration. FCP (Lu et al., 2023) proves that if the rank of quantile estimate $\hat{q}_\alpha$ is between $(1-\alpha-\epsilon)(N+K)$ and $(1-\alpha+\epsilon)(N+K)$, then we have:

$$F_1(\hat{q}_\alpha, \overline{\mathbf{v}}) \geq 1 - \alpha - \frac{\epsilon N_{\mathcal{B}} + 1}{N_{\mathcal{B}} + K_b}, \;\; F_2(\hat{q}_\alpha, \overline{\mathbf{v}}) \leq 1 - \alpha + \epsilon + \frac{K_b}{N_{\mathcal{B}} + K_b}. \tag{38}$$

Now we start deriving the error of $F_1(\cdot,\cdot)$ induced by the malicious clients. Let $\hat{\mathbf{v}} := \sum_{k\in\mathcal{B}} \dfrac{n_k}{N}\mathbf{v}^{(k)}$ be the estimated mean of characterization vector. Based on the results in part (b), we can derive as follows:

$$\left|F_1(\hat{q}_\alpha, \overline{\mathbf{v}}) - F_1(\hat{q}_\alpha, \hat{\mathbf{v}})\right| = \left|\sum_{j=1}^{H} \mathbb{I}\left[a_j < \hat{q}_\alpha\right] \overline{\mathbf{v}}_j - \sum_{j=1}^{H} \mathbb{I}\left[a_j < \hat{q}_\alpha\right] \hat{\mathbf{v}}_j\right| \tag{39}$$

$$\leq \sum_{j=1}^{H} \mathbb{I}\left[a_j < \hat{q}_\alpha\right] \left|\overline{\mathbf{v}}_j - \hat{\mathbf{v}}_j\right| \tag{40}$$

$$\leq \left\|\overline{\mathbf{v}} - \hat{\mathbf{v}}\right\|_1 \tag{41}$$

$$\leq \left(1 + \frac{N_m}{\min_{k'\in[K_b]n_{k'}}} \frac{2}{1-\tau}\right) r(\beta) \tag{42}$$

From triangular inequalities, we have:

$$F_1(\hat{q}_\alpha, \overline{\mathbf{v}}) - \left|F_1(\hat{q}_\alpha, \overline{\mathbf{v}}) - F_1(\hat{q}_\alpha, \hat{\mathbf{v}})\right| \leq F_1(\hat{q}_\alpha, \hat{\mathbf{v}}) \leq \mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right]. \tag{43}$$

Similarly, we can derive that $\left|F_2(\hat{q}_\alpha, \overline{\mathbf{v}}) - F_2(\hat{q}_\alpha, \hat{\mathbf{v}})\right| \leq \left(1 + \dfrac{N_m}{\min_{k'\in[K_b]n_{k'}}} \dfrac{2}{1-\tau}\right) r(\beta)$ and have:

$$F_2(\hat{q}_\alpha, \overline{\mathbf{v}}) + \left|F_2(\hat{q}_\alpha, \overline{\mathbf{v}}) - F_2(\hat{q}_\alpha, \hat{\mathbf{v}})\right| \geq F_2(\hat{q}_\alpha, \hat{\mathbf{v}}) \geq \mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right]. \tag{44}$$

Plugging in the terms in Equations (38) and (42) and leveraging the fact $N_{\mathcal{B}} \geq n_b$, we finally conclude that the following holds with probability $1-\beta$:

$$\begin{aligned}
&\mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] \geq 1 - \alpha - \frac{\epsilon n_b + 1}{n_b + K_b} - \frac{H\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{n_b}}\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right), \\
&\mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] \leq 1 - \alpha + \epsilon + \frac{K_b}{n_b + K_b} + \frac{H\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{n_b}}\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right).
\end{aligned} \tag{45}$$

where $\tau = K_m/K_b$ is the ratio of the number of malicious clients and the number of benign clients, $N_m := \sum_{k\in[K]\setminus[K_b]} n_k$ is the total sample size of malicious clients, $n_b := \min_{k'\in[K_b]} n_{k'}$ is the minimal sample size of benign clients, and $\Phi^{-1}(\cdot)$ denotes the inverse of the cumulative distribution function (CDF) of standard normal distribution.

□

Next, we start proving Theorem 1.

Theorem 3 (Restatement of Theorem 1)**.** *Consider FCP setting with K_b benign clients and K_m malicious clients. The k-th client reports the characterization vector $\mathbf{v}^{(k)}$ and local sample size n_k to the server. Assume that the benign characterization vector $\mathbf{v}^{(k)}$ is sampled from multinomial distribution $\mathcal{D}_k$ with the event probability $\overline{\mathbf{v}}^{(k)}$ for the k-th client ($k \in [K_b]$). We use σ to quantify the heterogeneity of benign vectors as $\sigma = \max_{k_1,k_2\in[K_b]} \|\overline{\mathbf{v}}^{(k_1)} - \overline{\mathbf{v}}^{(k_2)}\|_1$. Let ϵ be the data sketching error as Equation* (3)*. Under the assumption that $K_m < K_b$, the following holds for test instance (X_t, Y_t) with probability $1-\beta$:*

$$\begin{aligned}
&\mathbb{P}\Big[Y_t \in \hat{C}_\alpha(X_t)\Big] \geq 1-\alpha - P_{byz} - \frac{N_m\sigma}{n_b(1-\tau)} - \frac{\epsilon n_b + 1}{n_b + K_b} \\
&\mathbb{P}\Big[Y_t \in \hat{C}_\alpha(X_t)\Big] \leq 1-\alpha + P_{byz} + \frac{N_m\sigma}{n_b(1-\tau)} + \frac{\epsilon n_b + (\epsilon+1)K_b}{n_b + K_b} \\
&\textit{where} \quad P_{byz} = \frac{H\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{n_b}}\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right)
\end{aligned} \tag{46}$$

where $\tau = K_m/K_b$ is the ratio of the number of malicious clients and the number of benign clients, $N_m := \sum_{k\in[K]\setminus[K_b]} n_k$ is the total sample size of malicious clients, $n_b := \min_{k'\in[K_b]} n_{k'}$ is the minimal sample size of benign clients, and $\Phi^{-1}(\cdot)$ denotes the inverse of the cumulative distribution function (CDF) of standard normal distribution.

Proof. The general structure of the proof follows the proof of Lemma C.1. We will omit similar derivation and refer to the proof of Lemma C.1 for details. The proof consists of 3 parts: (a) concentration analysis of the characterization vectors $\mathbf{v}^{(k)}$ for benign clients ($1 \leq k \leq K_b$), (b) analysis of the algorithm of the identification of malicious clients, and (c) analysis of the error of the coverage bound.

Part (a): concentration analysis of the characterization vectors $\mathbf{v}^{(k)}$ for benign clients ($1 \leq k \leq K_b$).

Let $\overline{\mathbf{v}}^{(k)}$ be the event probability of the multinormial distribution $\mathcal{D}^{(k)}$ for $k \in [K_b]$. By applying binomial proportion approximate normal confidence interval and union bound as in Part (a) in the proof of Lemma C.1, with confidence $1-\beta$, we have:

$$\left\|\mathbf{v}^{(k)} - \overline{\mathbf{v}}^{(k)}\right\|_1 \leq r(\beta) := \frac{H\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{\min_{k'\in[K_b]} n_{k'}}}, \ \forall k \in [K_b], \tag{47}$$

where $r(\beta)$ is the perturbation radius of random vector $\mathbf{v}^{(k)}$ given confidence level $1-\beta$. $\forall k_1, k_2 \in [K_b]$, we can upper bound the ℓ_p norm distance between $\mathbf{v}^{(k_1)}$ and $\mathbf{v}^{(k_2)}$ as:

$$\begin{aligned}
\left\|\mathbf{v}^{(k_1)} - \mathbf{v}^{(k_2)}\right\|_p &\leq \left\|\mathbf{v}^{(k_1)} - \overline{\mathbf{v}}^{(k_1)}\right\|_p + \left\|\overline{\mathbf{v}}^{(k_1)} - \overline{\mathbf{v}}^{(k_2)}\right\|_p + \left\|\mathbf{v}^{(k_2)} - \overline{\mathbf{v}}^{(k_1)}\right\|_p && (48)\\
&\leq \left\|\mathbf{v}^{(k_1)} - \overline{\mathbf{v}}^{(k_1)}\right\|_1 + \left\|\overline{\mathbf{v}}^{(k_1)} - \overline{\mathbf{v}}^{(k_2)}\right\|_p + \left\|\mathbf{v}^{(k_2)} - \overline{\mathbf{v}}^{(k_1)}\right\|_1 && (49)\\
&\leq 2r(\beta) + \sigma, && (50)
\end{aligned}$$

where Equation (50) holds by Equation (47).

Part (b): analysis of the algorithm of the identification of malicious clients.

Let $N(k, n)$ be the set of the index of n nearest clients to the k-th client based on the metrics of ℓ_p norm distance in the space of characterization vectors. Then the maliciousness scores $M(k)$ for the k-th client ($k \in [K]$) can be defined as:

$$M(k) := \frac{1}{K_b - 1} \sum_{k'\in N(k,K_b-1)} \left\|\mathbf{v}^{(k)} - \mathbf{v}^{(k')}\right\|_p. \tag{51}$$

Let $\mathcal{B}$ be the set of the index of benign clients identified by Algorithm 1 by selecting the clients associated with the lowest K_b maliciousness scores. We will consider the following cases separately: (1) $\mathcal{B}$ contains exactly K_b benign clients, and (2) $\mathcal{B}$ contains at least one malicious client indexed by m.

Case (1): $\mathcal{B}$ ($|\mathcal{B}| = K_b$) contains exactly K_b benign clients. We can derive as follows:

$$\left\| \sum_{k=1}^{K_b} \frac{n_k}{N_b} \mathbf{v}^{(k)} - \sum_{k=1}^{K_b} \frac{n_k}{N_b} \overline{\mathbf{v}}^{(k)} \right\|_p \leq \sum_{k=1}^{K_b} \frac{n_k}{N_b} \left\| \mathbf{v}^{(k)} - \overline{\mathbf{v}}^{(k)} \right\|_p \tag{52}$$

$$\leq \sum_{k=1}^{K_b} \frac{n_k}{N_b} r(\beta) \tag{53}$$

$$= r(\beta), \tag{54}$$

where $N_b := \sum_{k \in [K_b]} n_k$ is the total sample size of benign clients.

Case (2): $\mathcal{B}$ ($|\mathcal{B}| = K_b$) contains at least one malicious client indexed by m. Since we assume $K_m < K_b$, there are at most $K_b - 1$ malicious clients in $\mathcal{B}$. Therefore, there is at least 1 benign client in $[K] \backslash \mathcal{B}$ indexed by b. From the fact that $M(m) \leq M(b)$ and expanding the definitions the maliciousness score as Part (b) in the proof of Lemma C.1, we get that $\exists b_b \in \mathcal{B}, b_b \in [K_b]$:

$$\left\| \mathbf{v}^{(m)} - \mathbf{v}^{(b_b)} \right\|_p \leq \frac{(K_b - 1)(2r(\beta) + \sigma)}{K_b - K_m} \tag{55}$$

Therefore, we can upper bound the distance between the estimated global event probability vector $\sum_{k \in \mathcal{B}} \frac{n_k}{N_\mathcal{B}} \mathbf{v}^{(k)}$ and the benign global event probability vector $\sum_{k \in [K_b]} \frac{n_k}{N_b} \overline{\mathbf{v}}^{(k)}$.

We first show that $\forall k \in [K_b]$, we have:

$$\left\| \mathbf{v}^{(k)} - \sum_{k \in [K_b]} \frac{n_k}{N_b} \overline{\mathbf{v}}^{(k)} \right\|_p \leq \sum_{k \in [K_b]} \frac{n_k}{N_b} \left\| \mathbf{v}^{(k)} - \overline{\mathbf{v}}^{(k)} \right\|_p \leq r(\beta). \tag{56}$$

Then, we can derive as follows:

$$\left\| \sum_{k \in \mathcal{B}} \frac{n_k}{N_\mathcal{B}} \mathbf{v}^{(k)} - \sum_{k \in [K_b]} \frac{n_k}{N_b} \overline{\mathbf{v}}^{(k)} \right\|_p \tag{57}$$

$$\leq \sum_{k \in \mathcal{B}, k \in [K_b]} \frac{n_k}{N_\mathcal{B}} \left\| \mathbf{v}^{(k)} - \sum_{k \in [K_b]} \frac{n_k}{N_b} \overline{\mathbf{v}}^{(k)} \right\|_p + \sum_{k \in \mathcal{B}, k \in [K] \backslash [K_b]} \frac{n_k}{N_\mathcal{B}} \left\| \mathbf{v}^{(k)} - \sum_{k \in [K_b]} \frac{n_k}{N_b} \overline{\mathbf{v}}^{(k)} \right\|_p \tag{58}$$

$$\leq \sum_{k \in \mathcal{B}, k \in [K_b]} \frac{n_k}{N_\mathcal{B}} r(\beta) + \sum_{k \in \mathcal{B}, k \in [K] \backslash [K_b]} \frac{n_k}{N_\mathcal{B}} \left[\left\| \mathbf{v}^{(k)} - \mathbf{v}^{(b_b)} \right\|_p + \left\| \mathbf{v}^{(b_b)} - \sum_{k \in [K_b]} \frac{n_k}{N_b} \overline{\mathbf{v}}^{(k)} \right\|_p \right] \tag{59}$$

$$\leq \sum_{k \in \mathcal{B}, k \in [K_b]} \frac{n_k}{N_\mathcal{B}} r(\beta) + \sum_{k \in \mathcal{B}, k \in [K] \backslash [K_b]} \frac{n_k}{N_\mathcal{B}} \left[\frac{(K_b - 1)(2r(\beta) + \sigma)}{K_b - K_m} + r(\beta) \right] \tag{60}$$

$$\leq r(\beta) + \sum_{k \in \mathcal{B}, k \in [K] \backslash [K_b]} \frac{n_k}{N_\mathcal{B}} \frac{(K_b - 1)(2r(\beta) + \sigma)}{K_b - K_m} \tag{61}$$

$$\leq r(\beta) \left(1 + \frac{N_m}{n_b} \frac{2}{1 - \tau} \right) + \frac{N_m}{n_b} \frac{\sigma}{1 - \tau}. \tag{62}$$

Part (c): analysis of the error of the coverage bound.

Let $F_1(q, \mathbf{v}) := \sum_{j=1}^{H} \mathbb{I}[a_j < q] \mathbf{v}_j$, where $q \in [0, 1]$ and a_j is the j-th partition point used to construct the characterization vector $\mathbf{v} \in \Delta^H$. Let $F_2(q, \mathbf{v}) := \sum_{j=1}^{H} \mathbb{I}[a_{j-1} < q] \mathbf{v}_j$. This part follows the same procedure to translate the error of aggregated vectors induced by malicious clients to the error of the bound of marginal coverage. The only difference is that considering data heterogeneity, the error of aggregated vectors formulated in Equation (62) needs additional correction by

the client data disparity. Therefore, by analyzing the connection between characterization vector and coverage similarly in Part (3) in the proof of Lemma C.1, we have:

$$|F_1(\hat{q}_\alpha, \overline{\mathbf{v}}) - F_1(\hat{q}_\alpha, \hat{\mathbf{v}})| \leq r(\beta)\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right) + \frac{N_m}{n_b}\frac{\sigma}{1-\tau}, \tag{63}$$

$$|F_2(\hat{q}_\alpha, \overline{\mathbf{v}}) - F_2(\hat{q}_\alpha, \hat{\mathbf{v}})| \leq r(\beta)\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right) + \frac{N_m}{n_b}\frac{\sigma}{1-\tau}, \tag{64}$$

where $\overline{\mathbf{v}} := \sum_{k\in[K_b]} \frac{n_k}{N_b}\overline{\mathbf{v}}^{(k)}$ and $\hat{\mathbf{v}} := \sum_{k\in\mathcal{B}} \frac{n_k}{N_\mathcal{B}}\mathbf{v}^{(k)}$. On the other hand, from triangular inequalities, we have:

$$F_1(\hat{q}_\alpha, \overline{\mathbf{v}}) - |F_1(\hat{q}_\alpha, \overline{\mathbf{v}}) - F_1(\hat{q}_\alpha, \hat{\mathbf{v}})| \leq F_1(\hat{q}_\alpha, \hat{\mathbf{v}}) \leq \mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right], \tag{65}$$

$$F_2(\hat{q}_\alpha, \overline{\mathbf{v}}) + |F_2(\hat{q}_\alpha, \overline{\mathbf{v}}) - F_2(\hat{q}_\alpha, \hat{\mathbf{v}})| \geq F_2(\hat{q}_\alpha, \hat{\mathbf{v}}) \geq \mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right]. \tag{66}$$

Plugging in the terms, we finally conclude that the following holds with probability $1-\beta$:

$$\begin{aligned}
&\mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] \geq 1 - \alpha - \frac{\epsilon n_b + 1}{n_b + K_b} - \frac{H\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{n_b}}\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right) - \frac{N_m}{n_b}\frac{\sigma}{1-\tau}, \\
&\mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] \leq 1 - \alpha + \epsilon + \frac{K_b}{n_b + K_b} + \frac{H\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{n_b}}\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right) + \frac{N_m}{n_b}\frac{\sigma}{1-\tau}.
\end{aligned} \tag{67}$$

where $\tau = K_m/K_b$ is the ratio of the number of malicious clients and the number of benign clients, $N_m := \sum_{k\in[K]\setminus[K_b]} n_k$ is the total sample size of malicious clients, $n_b := \min_{k'\in[K_b]} n_{k'}$ is the minimal sample size of benign clients, and $\Phi^{-1}(\cdot)$ denotes the inverse of CDF of the standard normal distribution.

□

C.2. Proof of Theorem 2

Theorem 4 (Restatement of Theorem 2)**.** *Assume $\mathbf{v}^{(k)}$ ($k \in [K_b]$) are IID sampled from Gaussian $\mathcal{N}(\mu, \Sigma)$ with mean $\mu \in \mathbb{R}^H$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{H\times H}$. Let $d := \min_{k\in[K]\setminus[K_b]} \|\mathbf{v}^{(k)} - \mu\|_2$. Suppose that we use ℓ_2 norm to measure vector distance and leverage the malicious client number estimator with an initial guess of a number of benign clients $\tilde{K}_b$ such that $K_m < \tilde{K}_b \leq K_b$. Then we have:*

$$\mathbb{P}\left[\hat{K}_m = K_m\right] \geq 1 - \frac{(3\tilde{K}_b - K_m - 2)^2 Tr(\Sigma)}{(\tilde{K}_b - K_m)^2 d^2} - \frac{2(K + K_b)Tr(\Sigma)\sigma_{max}^2(\Sigma^{-1/2})}{\sigma_{min}^2(\Sigma^{-1/2})d^2}, \tag{68}$$

where $\sigma_{max}(\Sigma^{-1/2})$, $\sigma_{min}(\Sigma^{-1/2})$ denote the maximal and minimal eigenvalue of matrix $\Sigma^{-1/2}$, and $Tr(\Sigma)$ denotes the trace of matrix Σ.

Proof. From the concentration inequality of multivariate Gaussian distribution (Vershynin, 2018), the following holds for $\mathbf{v}^{(k)} \sim \mathcal{N}(\mu, \Sigma)$:

$$\mathbb{P}\left[\|\mathbf{v}^{(k)} - \mu\|_2 \leq \sqrt{\frac{1}{\delta}\text{Tr}(\Sigma)}\right] \geq 1 - \delta. \tag{69}$$

Applying union bound for all benign clients $k \in [K_b]$, the following concentration bound holds:

$$\mathbb{P}\left[\|\mathbf{v}^{(k)} - \mu\|_2 \leq \sqrt{\frac{K_b}{\delta}\text{Tr}(\Sigma)},\ \forall k \in [K_b]\right] \geq 1 - \delta, \tag{70}$$

Let the perturbation radius $r := \dfrac{\tilde{K}_b - K_m}{3\tilde{K}_b - K_m - 2}d$. Then we can derive that:

$$\mathbb{P}\left[\|\mathbf{v}^{(k)} - \mu\|_2 \leq r := \frac{\tilde{K}_b - K_m}{3\tilde{K}_b - K_m - 2}d,\ \forall k \in [K_b]\right] \geq 1 - \frac{(3\tilde{K}_b - K_m - 2)^2\text{Tr}(\Sigma)}{(\tilde{K}_b - K_m)^2 d^2} := 1 - \delta. \tag{71}$$

The following discussion is based on the fact that $\|\mathbf{v}^{(k)} - \mu\|_2 \leq r := \frac{\tilde{K}_b - K_m}{3\tilde{K}_b - K_m - 2} d,\ \forall k \in [K_b]$, and the confidence $1 - \delta$ will be incorporated in the final statement. Let $N(k, n)$ be the index set of n nearest neighbors of client k in the characterization vector space with the metric of ℓ_2 norm distance. We consider the maliciousness score $M(b)$ of any benign client $b \in [K_b]$:

$$M(b) = \frac{1}{\tilde{K}_b - 1} \sum_{k' \in N(b, \tilde{K}_b - 1)} \left\| \mathbf{v}^{(b)} - \mathbf{v}^{(k')} \right\|_2 \tag{72}$$

$$\leq \max_{k' \in [K_b]} \left\| \mathbf{v}^{(b)} - \mathbf{v}^{(k')} \right\|_2 \tag{73}$$

$$\leq \max_{k' \in [K_b]} \left\{ \left\| \mathbf{v}^{(b)} - \mu \right\|_2 + \left\| \mu - \mathbf{v}^{(k')} \right\|_2 \right\} \tag{74}$$

$$\leq \frac{2(\tilde{K}_b - K_m)}{3\tilde{K}_b - K_m - 2} d. \tag{75}$$

Equation (73) holds since the average of distances to $\tilde{K}_b - 1$ nearest vectors is upper bounded by the average of distances to arbitrary $\tilde{K}_b - 1$ benign clients, which is upper bounded by the maximal distance to benign clients. Equation (75) holds by plugging in the results in Equation (71).

We consider the maliciousness score $M(m)$ of any malicious client $m \in [K] \backslash [K_b]$:

$$M(m) = \frac{1}{\tilde{K}_b - 1} \sum_{k' \in N(m, \tilde{K}_b - 1)} \left\| \mathbf{v}^{(m)} - \mathbf{v}^{(k')} \right\|_2 \tag{76}$$

$$\geq \frac{1}{\tilde{K}_b - 1} \sum_{k' \in N(m, \tilde{K}_b - 1), k' \in [K_b]} \left\| \mathbf{v}^{(m)} - \mathbf{v}^{(k')} \right\|_2 \tag{77}$$

$$\geq \frac{1}{\tilde{K}_b - 1} \sum_{k' \in N(m, \tilde{K}_b - 1), k' \in [K_b]} \left[\left\| \mathbf{v}^{(m)} - \mu \right\|_2 - \left\| \mu - \mathbf{v}^{(k')} \right\|_2 \right] \tag{78}$$

$$\geq \frac{1}{\tilde{K}_b - 1} (\tilde{K}_b - K_m) \left(d - \frac{\tilde{K}_b - K_m}{3\tilde{K}_b - K_m - 2} d \right) \tag{79}$$

$$\geq \frac{2(\tilde{K}_b - K_m)}{3\tilde{K}_b - K_m - 2} d. \tag{80}$$

Equation (79) holds since $d := \min_{k \in [K] \backslash [K_b]} \|\mathbf{v}^{(k)} - \mu\|_2$ by definition. Therefore, from Equation (75) and Equation (80), we can conclude that with probability $1 - \delta$, $M(m) \geq M(b),\ \forall b \in [K_b], m \in [K] \backslash [K_b]$, which implies that $\forall k \in [K_b], I(k) \in [K_b]$ and $\forall k \in [K] - [K_b], I(k) \in [K] \backslash [K_b]$.

Recall that the estimate of the number of benign clients $\hat{K}_b$ is given by:

$$\hat{K}_b = \arg\max_{z \in [K]} \left[\frac{1}{z} \sum_{k=1}^{z} \log p(\mathbf{v}^{(I(k))}; \mu, \Sigma) - \frac{1}{K - z} \sum_{k=z+1}^{K} \log p(\mathbf{v}^{(I(k))}; \mu, \Sigma) \right]. \tag{81}$$

For ease of notation, let $T(z) := \frac{1}{z} \sum_{k=1}^{z} \log p(\mathbf{v}^{(I(k))}; \mu, \Sigma) - \frac{1}{K - z} \sum_{k=z+1}^{K} \log p(\mathbf{v}^{(I(k))}; \mu, \Sigma)$ for $z \in [K]$ and $d_k := \mathbf{v}^{(I(k))} - \mu$ for $k \in [K]$. Then we can upper bound the probability of an underestimate of the number of malicious

clients $\mathbb{P}\left[\hat{K}_b < K_b\right]$ as follows:

$$
\begin{aligned}
&\mathbb{P}\left[\hat{K}_b < K_b\right] && (82)\\
=&\mathbb{P}\left[T(\hat{K}_b) > T(K_b)\right] && (83)\\
\leq&\mathbb{P}\left[\frac{-(K_b-\hat{K}_b)}{K_b\hat{K}_b}\sum_{k=1}^{\hat{K}_b}\log p(\mathbf{v}^{(I(k))}) + \frac{K-\hat{K}_b+K_b}{K_b(K-\hat{K}_b)}\sum_{k=\hat{K}_b+1}^{K_b}\log p(\mathbf{v}^{(I(k))})\right.\\
&\left. < \frac{K_b-\hat{K}_b}{(K-K_b)(K-\hat{K}_b)}\sum_{k=K_b+1}^{K}\log p(\mathbf{v}^{(I(k))})\right] && (84)\\
\leq&\mathbb{P}\left[\frac{K-\hat{K}_b+K_b}{K_b(K-\hat{K}_b)}\sum_{k=\hat{K}_b+1}^{K_b}-d_k^T\Sigma^{-1}d_k < \frac{K_b-\hat{K}_b}{(K-K_b)(K-\hat{K}_b)}\sum_{k=K_b+1}^{K}-d_k^T\Sigma^{-1}d_k\right] && (85)\\
\leq&\mathbb{P}\left[\frac{K_b-\hat{K}_b}{(K-K_b)}\sum_{k=K_b+1}^{K}\|d_k^T\Sigma^{-1/2}\|_2^2 < \frac{K-\hat{K}_b+K_b}{K_b}\sum_{k=\hat{K}_b+1}^{K_b}\|d_k^T\Sigma^{-1/2}\|_2^2\right] && (86)\\
\leq&\mathbb{P}\left[\frac{K_b-\hat{K}_b}{(K-K_b)}\sum_{k=K_b+1}^{K}\sigma_{\min}^2(\Sigma^{-1/2})\|d_k^T\|_2^2 < \frac{K-\hat{K}_b+K_b}{K_b}\sum_{k=\hat{K}_b+1}^{K_b}\sigma_{\max}^2(\Sigma^{-1/2})\|d_k^T\|_2^2\right] && (87)\\
\leq&\mathbb{P}\left[\sigma_{\min}^2(\Sigma^{-1/2})d^2 < \frac{K-\hat{K}_b+K_b}{K_b}\sigma_{\max}^2(\Sigma^{-1/2})\max_{k\in[K_b]}\|d_k^T\|_2^2\right] && (88)\\
\leq&\mathbb{P}\left[\max_{k\in[K_b]}\|d_k^T\|_2 > \sqrt{\frac{K_b}{K+K_b}}\frac{\sigma_{\min}(\Sigma^{-1/2})d}{\sigma_{\max}(\Sigma^{-1/2})}\right] && (89)\\
\leq&\frac{(K+K_b)\mathrm{Tr}(\Sigma)\sigma_{\max}^2(\Sigma^{-1/2})}{\sigma_{\min}^2(\Sigma^{-1/2})d^2} && (90)
\end{aligned}
$$

Equation (84) holds by plugging in the definitions in Equation (81) and rearranging the terms. Equation (85) holds by dropping the positive term $\frac{-(K_b-\hat{K}_b)}{K_b\hat{K}_b}\sum_{k=1}^{\hat{K}_b}\log p(\mathbf{v}^{(I(k))})$ and rearranging log-likelihood terms of multivariate Gaussian with d_k. Equation (87) holds by leveraging the fact that $\sigma_{\min}(\Sigma^{-1/2})\|d_k^T\|_2 \leq \|d_k^T\Sigma^{-1/2}\|_2 \leq \sigma_{\max}(\Sigma^{-1/2})\|d_k^T\|_2$. Similarly, we can upper bound the probability of overestimation of the number of malicious clients $\mathbb{P}\left[\hat{K}_b > K_b\right]$ as:

$$
\mathbb{P}\left[\hat{K}_b > K_b\right] \leq \frac{(K+K_b)\mathrm{Tr}(\Sigma)\sigma_{\max}^2(\Sigma^{-1/2})}{\sigma_{\min}^2(\Sigma^{-1/2})d^2}. \tag{91}
$$

We can finally conclude that:

$$
\mathbb{P}\left[\hat{K}_b = K_b\right] \geq 1 - \frac{(3\tilde{K}_b - K_m - 2)^2\mathrm{Tr}(\Sigma)}{(\tilde{K}_b - K_m)^2 d^2} - \frac{2(K+K_b)\mathrm{Tr}(\Sigma)\sigma_{\max}^2(\Sigma^{-1/2})}{\sigma_{\min}^2(\Sigma^{-1/2})d^2}. \tag{92}
$$

□

D. Improvements with DKW inequality

D.1. Improvement of Lemma C.1 with DKW inequality

Theorem 5 (Improvement of Lemma C.1)**.** *For K clients including K_b benign clients and $K_m := K - K_b$ malicious clients, each client reports a characterization vector $\mathbf{v}^{(k)} \in \Delta^H$ $(k \in [K])$ and a quantity $n_k \in \mathbb{Z}^+$ $(k \in [K])$ to the server. Suppose that the reported characterization vectors of benign clients are sampled from the same underlying multinomial distribution $\mathcal{D}$, while those of malicious clients can be arbitrary. Let ϵ be the estimation error of the data sketching by characterization vectors as illustrated in Equation* (3)*. Under the assumption that $K_m < K_b$, the following holds with probability $1-\beta$:*

$$\begin{aligned}
&\mathbb{P}\left[Y_{test} \in \hat{C}_\alpha(X_{test})\right] \geq 1 - \alpha - \frac{\epsilon n_b + 1}{n_b + K_b} - H\sqrt{\frac{\ln(2K_b/\beta)}{2n_b}}\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right), \\
&\mathbb{P}\left[Y_{test} \in \hat{C}_\alpha(X_{test})\right] \leq 1 - \alpha + \epsilon + \frac{K_b}{n_b + K_b} + H\sqrt{\frac{\ln(2K_b/\beta)}{2n_b}}\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right),
\end{aligned} \tag{93}$$

where $\tau = K_m/K_b$ is the ratio of the number of malicious clients and the number of benign clients, $N_m := \sum_{k\in[K]\setminus[K_b]} n_k$ is the total sample size of malicious clients, and $n_b := \min_{k'\in[K_b]} n_{k'}$ is the minimal sample size of benign clients.

Proof. The proof structure follows the proof of Lemma C.1 and consists of 3 parts: (a) concentration analysis of the characterization vectors $\mathbf{v}^{(k)}$ for benign clients $(1 \leq k \leq K_b)$, (b) analysis of the algorithm of the identification of malicious clients, and (c) analysis of the error of the coverage bound. Part (b) and (c) are exactly the same as the proof Lemma C.1 and the only difference lies in the use of a more advanced concentration bound in part (a), which provides concentration analysis of the characterization vectors $\mathbf{v}^{(k)}$ for benign clients $(1 \leq k \leq K_b)$. Let $\mathbf{v}_h^{(k)}$ be the h-th element of vector $\mathbf{v}^{(k)}$. According to the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, we have:

$$\mathbb{P}\left[\left|\mathbf{v}_h^{(k)} - \overline{\mathbf{v}}_h\right| > \beta\right] \leq 2\exp\left\{-2H\beta^2\right\}, \ \forall h \in \{1, 2, .., H\}. \tag{94}$$

Applying the union bound for K_b characterization vectors of benign clients, the following holds with probability $1-\beta$:

$$\left|\mathbf{v}_h^{(k)} - \overline{\mathbf{v}}_h\right| \leq \sqrt{\frac{\ln(2K_b/\beta)}{2n_b}}, \ \forall k \in [K_b], \ \forall h \in [H], \tag{95}$$

from which we can derive the bound of difference for ℓ_1 norm distance as:

$$\left\|\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_1 \leq r(\beta) := H\sqrt{\frac{\ln(2K_b/\beta)}{2n_b}}, \ \forall k \in [K_b], \tag{96}$$

where $r(\beta)$ is the perturbation radius of random vector $\mathbf{v}$ given confidence level $1-\beta$. $\forall k_1, k_2 \in [K_b]$, the following holds with probability $1-\beta$ due to the triangular inequality:

$$\left\|\mathbf{v}^{(k_1)} - \mathbf{v}^{(k_2)}\right\|_1 \leq \left\|\mathbf{v}^{(k_1)} - \overline{\mathbf{v}}\right\|_1 + \left\|\mathbf{v}^{(k_2)} - \overline{\mathbf{v}}\right\|_1 \leq 2r(\beta). \tag{97}$$

Furthermore, due to the fact that $\|\mathbf{v}\|_p \leq \|\mathbf{v}\|_1$ for any integer $p \geq 1$, the following holds with probability $1-\beta$:

$$\left\|\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_p \leq \left\|\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_1 \leq r(\beta), \tag{98}$$

$$\left\|\mathbf{v}^{(k_1)} - \mathbf{v}^{(k_2)}\right\|_p \leq \left\|\mathbf{v}^{(k_1)} - \mathbf{v}^{(k_2)}\right\|_1 \leq 2r(\beta). \tag{99}$$

Then following the part (b) and (c) in the proof of Lemma C.1, we can finally conclude that:

$$\begin{aligned}
&\mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] \geq 1 - \alpha - \frac{\epsilon n_b + 1}{n_b + K_b} - H\sqrt{\frac{\ln(2K_b/\beta)}{2n_b}}\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right), \\
&\mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] \leq 1 - \alpha + \epsilon + \frac{K_b}{n_b + K_b} + H\sqrt{\frac{\ln(2K_b/\beta)}{2n_b}}\left(1 + \frac{N_m}{n_b}\frac{2}{1-\tau}\right),
\end{aligned} \tag{100}$$

□

D.2. Improvement of Theorem 1 with DKW inequality

Theorem 6 (Improvement of Theorem 1 with DKW inequality)**.** *Under the same definitions and conditions in Lemma C.1, the following holds with probability* $1-\beta$*:*

$$\begin{aligned}
\mathbb{P}\left[Y_{test} \in \hat{C}_\alpha(X_{test})\right] &\geq 1-\alpha-\frac{\epsilon n_b+1}{n_b+K_b}-H\sqrt{\frac{\ln(2K_b/\beta)}{2n_b}}\left(1+\frac{N_m}{n_b}\frac{2}{1-\tau}\right)-\frac{N_m}{n_b}\frac{\sigma}{1-\tau}, \\
\mathbb{P}\left[Y_{test} \in \hat{C}_\alpha(X_{test})\right] &\leq 1-\alpha+\epsilon+\frac{K_b}{n_b+K_b}+H\sqrt{\frac{\ln(2K_b/\beta)}{2n_b}}\left(1+\frac{N_m}{n_b}\frac{2}{1-\tau}\right)+\frac{N_m}{n_b}\frac{\sigma}{1-\tau}.
\end{aligned} \tag{101}$$

Proof. We conclude the proof by leveraging the concentration analysis in the proof of Theorem 5 and part (b) and part (c) in the proof of Theorem 1. □

E. Analysis of Rob-FCP with an overestimated number of benign clients K_b'

Theorem 7 (Lemma C.1 with an overestimated number of benign clients)**.** *For K clients including K_b benign clients and $K_m := K - K_b$ malicious clients, each client reports a characterization vector $\mathbf{v}^{(k)} \in \Delta^H$ $(k \in [K])$ and a quantity $n_k \in \mathbb{Z}^+$ $(k \in [K])$ to the server. Suppose that the reported characterization vectors of benign clients are sampled from the same underlying multinomial distribution $\mathcal{D}$, while those of malicious clients can be arbitrary. Let ϵ be the estimation error of the data sketching by characterization vectors as illustrated in Equation* (3)*. **Let** $K_b' > K_b$ **be the overestimated number of benign clients.** We also assume benign clients and malicious clients have the same sample sizes. Under the assumption that $K_m < K_b$, the following holds with probability $1-\beta$:*

$$\begin{aligned}
\mathbb{P}\left[Y_{test} \in \hat{C}_\alpha(X_{test})\right] &\geq 1-\alpha-\frac{\epsilon n_b+1}{n_b+K_b}-\left[1-\frac{K_b}{K_b'}\left(1-\frac{H\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{n_b}}\right)\right], \\
\mathbb{P}\left[Y_{test} \in \hat{C}_\alpha(X_{test})\right] &\leq 1-\alpha+\epsilon+\frac{K_b}{n_b+K_b}+\left[1-\frac{K_b}{K_b'}\left(1-\frac{H\Phi^{-1}(1-\beta/2HK_b)}{2\sqrt{n_b}}\right)\right],
\end{aligned} \tag{102}$$

where $\tau = K_m/K_b$ is the ratio of the number of malicious clients and the number of benign clients, $N_m := \sum_{k\in[K]\setminus[K_b]} n_k$ is the total sample size of malicious clients, $n_b := \min_{k'\in[K_b]} n_{k'}$ is the minimal sample size of benign clients, and $\Phi^{-1}(\cdot)$ denotes the inverse of the cumulative distribution function (CDF) of standard normal distribution.

Proof. The proof consists of 3 parts: (a) concentration analysis of the characterization vectors $\mathbf{v}^{(k)}$ for benign clients $(1 \leq k \leq K_b)$, (b) analysis of the algorithm of the identification of malicious clients, and (c) analysis of the error of the coverage bound. Part (a) and (c) follow that of Lemma C.1, and thus, we provide the details of part (b) here. Let $N(k,n)$ be the set of the index of n nearest clients to the k-th client based on the metrics of ℓ_p norm distance in the space of characterization vectors. Then the maliciousness scores $M(k)$ for the k-th client $(k \in [K])$ can be defined as:

$$M(k) := \frac{1}{K_b-1}\sum_{k'\in N(k,K_b-1)}\left\|\mathbf{v}^{(k)}-\mathbf{v}^{(k')}\right\|_p. \tag{103}$$

Let $\mathcal{B}$ be the set of the index of benign clients identified by Algorithm 1 by selecting the clients associated with the lowest K_b' maliciousness scores. We will consider the following cases separately: (1) $\mathcal{B}$ contains exactly K_b benign clients, and (2) $\mathcal{B}$ contains at least one malicious client indexed by m. *Case (1)*: $\mathcal{B}$ $(|\mathcal{B}| = K_b')$ contains all K_b benign clients. We can

derive as follows:

$$\left\|\sum_{k\in\mathcal{B}}\frac{n_k}{N_{\mathcal{B}}}\mathbf{v}^{(k)}-\overline{\mathbf{v}}\right\|_p \leq \sum_{k\in\mathcal{B}}\frac{n_k}{N_{\mathcal{B}}}\left\|\mathbf{v}^{(k)}-\overline{\mathbf{v}}\right\|_p \tag{104}$$

$$\leq \sum_{k\in\mathcal{B},k\in[K_b]}\frac{n_k}{N_{\mathcal{B}}}\left\|\mathbf{v}^{(k)}-\overline{\mathbf{v}}\right\|_p + \sum_{k\in\mathcal{B},k\in[K]\setminus[K_b]}\frac{n_k}{N_{\mathcal{B}}}\left\|\mathbf{v}^{(k)}-\overline{\mathbf{v}}\right\|_p \tag{105}$$

$$\leq \sum_{k\in\mathcal{B},k\in[K_b]}\frac{n_k}{N_{\mathcal{B}}}r(\beta) + \sum_{k\in\mathcal{B},k\in[K]\setminus[K_b]}\frac{n_k}{N_{\mathcal{B}}}\times 1 \tag{106}$$

$$= \frac{K_b}{K_b'}r(\beta) + \left(1-\frac{K_b}{K_b'}\right) \tag{107}$$

$$= 1-\frac{K_b}{K_b'}\left(1-r(\beta)\right) \tag{108}$$

Case (2): $\mathcal{B}$ ($|\mathcal{B}| = K_b'$) does not contain all benign clients, which implicates that for any malicious client $m \in \mathcal{B}$, we can derive the lower bound of the maliciousness score for the m-th client $M(m)$ as:

$$M(m) = \frac{1}{K_b'-1}\sum_{k'\in N(m,K_b'-1)}\left\|\mathbf{v}^{(m)}-\mathbf{v}^{(k')}\right\|_p \tag{109}$$

$$\geq \frac{1}{K_b'-1}\sum_{k'\in N(m,K_b'-1),k'\in[K_b]}\left\|\mathbf{v}^{(m)}-\mathbf{v}^{(k')}\right\|_p. \tag{110}$$

Since there are at least $K_b' - K_m$ benign clients in $\mathcal{B}$ (there are at most K_m malicious clients in $\mathcal{B}$), there exists one client indexed by b_b ($b_b \in \mathcal{B}$) such that:

$$\left\|\mathbf{v}^{(m)}-\mathbf{v}^{(b_b)}\right\|_p \leq \frac{(K_b'-1)M(m)}{K_b'-K_m} \tag{111}$$

We can derive the upper bound of the maliciousness score for the b-th benign client (one benign client not in $\mathcal{B}$) $M(b)$ as:

$$M(b) = \frac{1}{K_b'-1}\sum_{k'\in N(b,K_b'-1)}\left\|\mathbf{v}^{(b)}-\mathbf{v}^{(k')}\right\|_p \tag{112}$$

$$\leq \frac{K_b-1}{K_b'-1}2r(\beta) + \frac{K_b-K_b'}{K_b'-1} \tag{113}$$

Since the m-th client is included in $\mathcal{B}$ and identified as a benign client, while the b-th client is not in $\mathcal{B}$, the following holds according to the procedure in Algorithm 1:

$$M(b) \geq M(m), \tag{114}$$

Then, we can derive the upper bound of $\left\|\mathbf{v}^{(m)}-\overline{\mathbf{v}}\right\|_p$, $\forall m \in \mathcal{B}$ and $K_b < m \leq K$ as follows:

$$\left\|\mathbf{v}^{(m)}-\overline{\mathbf{v}}\right\|_p \leq \left\|\mathbf{v}^{(m)}-\mathbf{v}^{(b_b)}\right\|_p + \left\|\mathbf{v}^{(b_b)}-\overline{\mathbf{v}}\right\|_p \tag{115}$$

$$\leq \frac{(K_b-1)2r(\beta)+K_b-K_b'}{K_b'-K_m} \tag{116}$$

Finally, we can derive as follows:

$$\left\|\sum_{k\in\mathcal{B}}\frac{n_k}{N_{\mathcal{B}}}\mathbf{v}^{(k)}-\overline{\mathbf{v}}\right\|_p \leq \sum_{k\in\mathcal{B}}\frac{n_k}{N_{\mathcal{B}}}\left\|\mathbf{v}^{(k)}-\overline{\mathbf{v}}\right\|_p \tag{117}$$

$$\leq \sum_{k\in\mathcal{B},k\in[K_b]}\frac{n_k}{N_{\mathcal{B}}}\left\|\mathbf{v}^{(k)}-\overline{\mathbf{v}}\right\|_p + \sum_{k\in\mathcal{B},k\in[K]\setminus[K_b]}\frac{n_k}{N_{\mathcal{B}}}\left\|\mathbf{v}^{(k)}-\overline{\mathbf{v}}\right\|_p \tag{118}$$

$$\leq \sum_{k\in\mathcal{B},k\in[K_b]}\frac{n_k}{N_{\mathcal{B}}}r(\beta) + \sum_{k\in\mathcal{B},k\in[K]\setminus[K_b]}\frac{n_k}{N_{\mathcal{B}}}\frac{(K_b-1)2r(\beta)+K_b-K_b'}{K_b'-K_m} \tag{119}$$

$$\leq \frac{K_b}{K_b'}r(\beta) + \frac{K_b'-K_b}{K_b'}\frac{(K_b-1)2r(\beta)+K_b-K_b'}{K_b'-K_m} \tag{120}$$

Algorithm 1 Malicious client identification

Input: number of clients K, number of benign clients K_b, sets of scores for K clients $\left\{s_j^{(k)}\right\}_{j\in[n_k],k\in[K]}$, parameter p in ℓ_p norm distance.
Output: set of benign clients $\mathcal{B}_{\text{Rob-FCP}}$.
for $k = 1$ **to** K **do**
  Characterize the conformity score observations $\left\{s_j^{(k)}\right\}_{j\in[n_k]}$ with a vector $\mathbf{v}^{(k)}$ for client k as Equation (4).
end for
for $k_1 = 1$ **to** K **do**
  for $k_2 = 1$ **to** K **do**
    Compute the vector distance $d_{k_1,k_2} \leftarrow \|\mathbf{v}^{(k_1)} - \mathbf{v}^{(k_2)}\|_p$.
  end for
end for
for $k = 1$ **to** K **do**
  Compute the set of index of $K_b - 1$ nearest neighbors for client k: $N_{ear}(k, K_b - 1)$.
  Compute maliciousness scores of client k as $M(k) \leftarrow \frac{1}{K_b - 1}\sum_{k'\in N_{ear}(k,K_b-1)} d_{k,k'}$.
end for
Compute the index set of benign clients $\mathcal{B}_{\text{Rob-FCP}}$ as the associated index of the lowest K_b maliciousness scores in $\{M(k))\}_{k=1}^K$.

Combining *case (1)* and *case (2)*, we can conclude that:

$$\begin{aligned}\left\|\sum_{k\in\mathcal{B}}\frac{n_k}{N_{\mathcal{B}}}\mathbf{v}^{(k)} - \overline{\mathbf{v}}\right\|_p &\leq \max\left\{1 - \frac{K_b}{K_b'}\left(1 - r(\beta)\right), \frac{K_b}{K_b'}r(\beta) + \frac{K_b' - K_b}{K_b'}\frac{(K_b-1)2r(\beta) + K_b - K_b'}{K_b' - K_m}\right\} \\ &= 1 - \frac{K_b}{K_b'}\left(1 - r(\beta)\right)\end{aligned} \tag{121}$$

Finally, by applying the analysis of part (a) and (c) in the proof of Lemma C.1, we can conclude that:

$$\begin{aligned}\mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] &\geq 1 - \alpha - \frac{\epsilon n_b + 1}{n_b + K_b} - \left[1 - \frac{K_b}{K_b'}\left(1 - \frac{H\Phi^{-1}(1 - \beta/2HK_b)}{2\sqrt{n_b}}\right)\right], \\ \mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] &\leq 1 - \alpha + \epsilon + \frac{K_b}{n_b + K_b} + \left[1 - \frac{K_b}{K_b'}\left(1 - \frac{H\Phi^{-1}(1 - \beta/2HK_b)}{2\sqrt{n_b}}\right)\right],\end{aligned} \tag{122}$$

□

F. Algorithm of Rob-FCP

We provide the complete pseudocodes of malicious client identification in Rob-FCP in Algorithm 1. First, we characterize the conformity scores $\{s_j^{(k)}\}_{j\in[n_k]}$ with a vector $\mathbf{v}^{(k)} \in \mathbb{R}^H$ for client k ($k \in [K]$) via histogram statistics as Equation (4). Then, we compute the pairwise ℓ_p-norm ($p \in \mathbb{Z}^+$) vector distance and the maliciousness scores for clients, which are the averaged vector distance to the clients in the $K_b - 1$ nearest neighbors, where K_b is the number of benign clients. Finally, the benign set identified by Rob-FCP $\mathcal{B}_{\text{Rob-FCP}}$ is the set of the index of the clients with the lowest K_b maliciousness scores in $\{M(k)\}_{k=1}^K$.

G. Experiments

G.1. Experiment setup

Datasets. We evaluate Rob-FCP on computer vision datasets including MNIST (Deng, 2012), CIFAR-10 (Krizhevsky et al.), and Tiny-ImageNet (T-ImageNet) (Le & Yang, 2015). We additionally evaluate Rob-FCP on two realistic healthcare datasets, including SHHS (Zhang et al., 2018) and PathMNIST (Yang et al., 2023). The MNIST dataset consists of a

Table 3: Benign conformal prediction results (marginal coverage / average set size) without any malicious clients.

Data Partition	$\beta = 0.0$	$\beta = 0.5$
MNIST	0.898 / 0.900	0.902 / 1.828
CIFAR-10	0.901 / 1.597	0.898 / 2.308
Tiny-ImageNet	0.901 / 21.92	0.899 / 42.35
SHHS	0.898 / 1.352	0.897 / 1.351
PathMNIST	0.904 / 1.242	0.901 / 1.361

collection of 70,000 handwritten digit images, each of which is labeled with the corresponding digit (0 through 9) that the image represents. CIFAR-10 consists of 60,000 32x32 color images, each belonging to one of the following 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Tiny-ImageNet consists of 200 different classes, each represented by 500 training images, making a total of 100,000 training images. Additionally, it has 10,000 validation images and 10,000 test images, with 50 images per class for both validation and test sets. Each image in Tiny-ImageNet is a 64x64 colored image. SHHS (the Sleep Heart Health Study) is a large-scale multi-center study to determine consequences of sleep-disordered breathing. We use the EEG recordings from SHHS for the sleep-staging task, where every 30-second-epoch is classified into Wake, N1, N2, N3 and REM stages. 2,514 patients (2,545,869 samples) were used for training the DNN, and 2,514 patients (2,543,550 samples) were used for calibration and testing. PathMNIST is a 9-class classification dataset consisting of 107,180 hematoxylin and eosin stained histological images. 89,996 images were used to train the DNN and 7,180 were used for calibration and testing.

Training and evaluation strategy. Except for SHHS, we partition the datasets by sampling the proportion of each label from Dirichlet distribution parameterized by β for every agent, following the literature (Li et al., 2022a). For SHHS, we assign the patients to different clients according to the proportion of their time being awake. The parameter of the Dirichlet distribution is fixed as 0.5 across the evaluations. We pretrain the models with standard FedAvg algorithm (McMahan et al., 2016). We use the same collaboratively pretrained model for conformal prediction for different methods for fair comparisons. We perform conformal prediction with nonconformity scores LAC (Sadinle et al., 2019) and APS (Romano et al., 2020). Without specification, we use the LAC score by default across evaluations. Given a pretrained estimator $\hat{\pi} : \mathbb{R}^d \mapsto \Delta^C$ with d-dimensional input and C classes, the LAC non-conformity score is formulated as:

$$S^{\text{LAC}}_{\hat{\pi}_y}(x, y) = 1 - \hat{\pi}_y(x). \tag{123}$$

The APS non-conformity score is formulated as:

$$S^{\text{APS}}_{\hat{\pi}_y}(x, y) = \sum\nolimits_{j \in \mathcal{Y}} \hat{\pi}_j(x) \mathbb{I}[\hat{\pi}_j(x) > \hat{\pi}_y(x)] + \hat{\pi}_y(x) u, \tag{124}$$

where $\mathbb{I}[\cdot]$ is the indicator function and u is uniformly sampled over the interval $[0, 1]$.

Byzantine attacks. To evaluate the robustness of Rob-FCP in the Byzantine setting, we compare Rob-FCP with the baseline FCP (Lu et al., 2023) under three types of Byzantine attacks: (1) *coverage attack* (CovAttack) which reports the largest conformity scores to induce a larger conformity score at the desired quantile and a lower coverage accordingly, (2) *efficiency attack* (EffAttack) which reports the smallest conformity scores to induce a lower conformity score at the quantile and a larger prediction set, and (3) Gaussian Attack (GauAttack) which injects random Gaussian noises to the scores to perturb the conformal calibration. The gaussian noises are sampled from a univariate Gaussian $\mathcal{N}(0, 0.5)$ with zero mean and 0.5 variance.

G.2. Additional evaluation results

Robustness of Rob-FCP across varying levels of data heterogeneity Data heterogeneity among clients poses significant challenges to achieving precise federated conformal prediction. To assess the resilience of Rob-FCP to this issue, we conducted evaluations using various values of the Dirichlet parameter β, which modulates the degree of data heterogeneity among clients. The results in Table 2 show that Rob-FCP reliably maintains marginal coverage and average set size at levels close to those anticipated, underscoring its robustness in the face of data skewness. Furthermore, we investigate additional approaches to create heterogeneous data that mirror demographic differences. This involves dividing the SHHS

Table 4: Marginal coverage / average set size on SHHS with heterogeneous data partition based on different attributes: wake time, N1, N2, N3, REM. The evaluation is done under different Coverage attack with 40% ($K_m/K = 40\%$) malicious clients. The desired marginal coverage is 0.9.

	wake time	N1	N2	N3	REM
FCP (SHHS)	0.835 / 1.098	0.841 / 1.104	0.841 / 1.104	0.837 / 1.105	0.840 / 1.107
Rob-FCP (SHHS)	0.901 / 1.367	0.902 / 1.358	0.902 / 1.355	0.902 / 1.375	0.900 / 1.356

Table 5: Runtime of RobFCP quantile computation with 40% malicious clients. The valuation is done on a RTX A6000 GPU.

	MNIST	CIFAR-10	Tiny-ImageNet	SHHS	PathMNIST
Runtime (seconds)	0.5284	0.5169	0.5563	0.2227	0.3032

dataset according to five specific attributes (wake time, N1, N2, N3, REM) and allocating instances to clients based on varying intervals of these attributes. The findings, detailed in Table 4, highlight Rob-FCP's capability to effectively handle diverse forms of data heterogeneity.

Runtime of Rob-FCP We evaluate the runtime of quantile computation in Rob-FCP in Table 5, which indicates the efficiency of federated conformal prediction with Rob-FCP.

Results with an overestimate or underestimate of the number of malicious clients In Table 6, we provided evaluations of Rob-FCP with incorrect numbers of malicious clients. The results show that either underestimated numbers or overestimated numbers would harm the performance to different extents. Specifically, an underestimate of the number of malicious clients will definitely lead to the inclusion of malicious clients in the identified set $\mathcal{B}$ and downgrade the quality of conformal prediction. On the other hand, an overestimated number will lead to the exclusion of some benign clients. The neglect of non-conformity scores of those clients will lead to a distribution shift from the true data distribution in the calibration process, breaking the data exchangeability assumption of conformal prediction, and a downgraded performance. Therefore, correctly estimating the number of malicious clients is of significance, and this is why we propose the malicious client number estimator, which is sound both theoretically and empirically to achieve the goal.

Benign conformal performance The benign conformal prediction performance (marginal coverage / average set size) without any malicious clients is provided in Table 3. As expected, the coverage of the prediction sets is very close to the target (0.9). In the setting with data heterogeneity across clients (i.e., $\beta = 0.5$), the predictive performance of the base global model is typically worse, leading to a larger average size of the prediction sets.

Byzantine robustness of Rob-FCP with known K_m We evaluate the marginal coverage and average set size of Rob-FCP under coverage, efficiency, and Gaussian attack and compare the results with the baseline FCP. We present results of FCP and Rob-FCP in existence of $10\%, 20\%, 30\%$ ($K_m/K = 10\%, 20\%, 30\%$) malicious clients on MNIST, CIFAR-10, Tiny-ImageNet (T-ImageNet), SHHS, and PathMNIST in Table 9. The coverage of FCP deviates drastically from the desired coverage level 0.9 under Byzantine attacks, along with a deviation from the benign set size. In contrast, Rob-FCP achieves comparable marginal coverage and average set size to the benign conformal performance.

Byzantine robustness of Rob-FCP with unknown K_m Similar to above, we evaluate the marginal coverage and average set size of Rob-FCP under verious attacks and compare the results with the FCP. We present results in existence of $10\%, 20\%, 30\%, 40\%$ ($K_m/K = 10\%, 20\%, 30\%, 40\%$) malicious clients in Table 10, where the number of the malicious clients is unknown to the algorithm. Again, the coverage of FCP as well as the size of the prediction set deviates drastically from the benign set setting, but Rob-FCP achieves comparable marginal coverage and average set size to the benign performance.

Robustness of Rob-FCP against mimic attacks We also evaluate the performance of Rob-FCP against the mimic attack strategy (Karimireddy et al., 2022), wherein malicious clients replicate the score statistics of a randomly chosen benign

Table 6: Marginal coverage / average set size under different Coverage attack with underestimated and overestimated numbers of malicious clients on TinyImageNet. The true ratio of malicious clients is 40% ($K_m/K = 25\%$), while we evaluate Rob-FCP with different ratios of malicious clients K'_m/K ranging from 5% to 45%. The desired marginal coverage is 0.9.

K'_m/K	5%	10%	15%	20%	25%	30%	35%	40%	45%
Coverage	0.8682	0.8756	0.8812	0.8884	0.9078	0.8936	0.8921/	0.8834	0.8803
Set Size	35.875	37.130	38.372	40.643	44.578	42.173	42.023	38.346	38.023

Table 7: Marginal coverage / average set size on Tiny-ImageNet with the desired level 0.9. The evaluation is conducted with different ratios of malicious clients K_m/K and different degrees of data heterogeneity β under mimic attack (MA). Mimic attack can not effectively distort the coverage for different data heterogeneity; Rob-FCP also maintains the coverage robustly.

	$\beta = 0.0$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 0.9$
Benign	0.898 / 21.728	0.896 / 43.038	0.903 / 43.864	0.899 / 42.352	0.902 / 43.843	0.904 / 43.919
MA ($K_m/K = 10\%$)	0.900 / 22.251	0.904 / 44.684	0.905 / 44.701	0.891 / 42.277	0.898 / 42.939	0.906 / 44.240
MA + Rob-FCP ($K_m/K = 10\%$)	0.903 / 23.823	0.893 / 41.994	0.899 / 43.169	0.901 / 43.734	0.909 / 44.811	0.897 / 42.632
MA ($K_m/K = 20\%$)	0.895 / 22.243	0.894 / 42.738	0.894 / 42.412	0.893 / 41.633	0.904 / 44.878	0.906 / 43.941
MA + Rob-FCP ($K_m/K = 20\%$)	0.902 / 22.651	0.895 / 41.575	0.901 / 42.770	0.905 / 44.124	0.899 / 43.793	0.897 / 42.589
MA ($K_m/K = 30\%$)	0.905 / 23.414	0.910 / 46.940	0.883 / 37.411	0.899 / 42.766	0.888 / 41.012	0.896 / 41.846
MA + Rob-FCP ($K_m/K = 30\%$)	0.898 / 22.839	0.912 / 47.390	0.906 / 44.882	0.893 / 41.481	0.906 / 45.372	0.897 / 42.813
MA ($K_m/K = 40\%$)	0.892 / 19.629	0.901 / 44.468	0.896 / 42.526	0.908 / 46.017	0.911 / 46.445	0.914 / 47.553
MA + Rob-FCP ($K_m/K = 40\%$)	0.899 / 20.952	0.904 / 45.023	0.905 / 43.368	0.908 / 46.518	0.915 / 47.023	0.892 / 40.561

client. It's critical to note that such strategies presuppose that the attackers have knowledge of the benign clients' score statistics, implying a more restricted threat model. We conduct the evaluations on Tiny-ImageNet with $1 - \alpha = 0.9$ with different ratios of malicious clients K_m/K. The results in Table 7 show that (1) across various degrees of data heterogeneity, merely approximating the scores of benign clients is insufficient to significantly impair the performance of conformal prediction; and (2) Rob-FCP still maintains the desired coverage under such attacks.

Robustness of Rob-FCP with different conformity scores Besides applying LAC nonconformity scores, we also evaluate Rob-FCP with APS scores (Romano et al., 2020). The results in Figures 6 to 11 demonstrate the Byzantine robustness of Rob-FCP with APS scores.

Ablation study of different conformity score distribution characterization One key step in Rob-FCP is to characterize the conformity score distribution based on empirical observations. We adopt the histogram statistics approach as Equation (4). Rob-FCP also flexibly allows for alternative approaches to characterizing the empirical conformity score samples with a real-valued vector $\mathbf{v}$. We can fit a parametric model (e.g., Gaussian model) to the empirical scores and concatenate the parameters as the characterization vector $\mathbf{v}$. Another alternative is to characterize the score samples with exemplars approximated by clustering algorithms such as KMeans. We empirically compare different approaches in Figure 12 and show that the histogram statistics approach achieves the best performance.

Ablation study of the distance measurement In Rob-FCP, we need to compute the distance between characterization vectors with measurement $d(\cdot, \cdot)$. We evaluate Rob-FCP with ℓ_1, ℓ_2, ℓ_∞-norm based vector distance as Equation (5) and an alternative Cosine similarity in Figure 13. The results show that the effectiveness of Rob-FCP is agnostic to these commonly used distance measurements. We adopt ℓ_2-norm vector distance for consistency across evaluations.

Table 8: Marginal coverage / average set size of Rob-FCP on Tiny-ImageNet with the desired level 0.9 under Gaussian Attack with standard deviation 0.5 inexistence of 40% malicious clients.

H	2	10	100	1000	10000
Marginal coverage / average set size	0.718 / 12.293	0.888 / 40.250	0.901 / 43.349	0.907 / 44.677	0.803 / 26.343

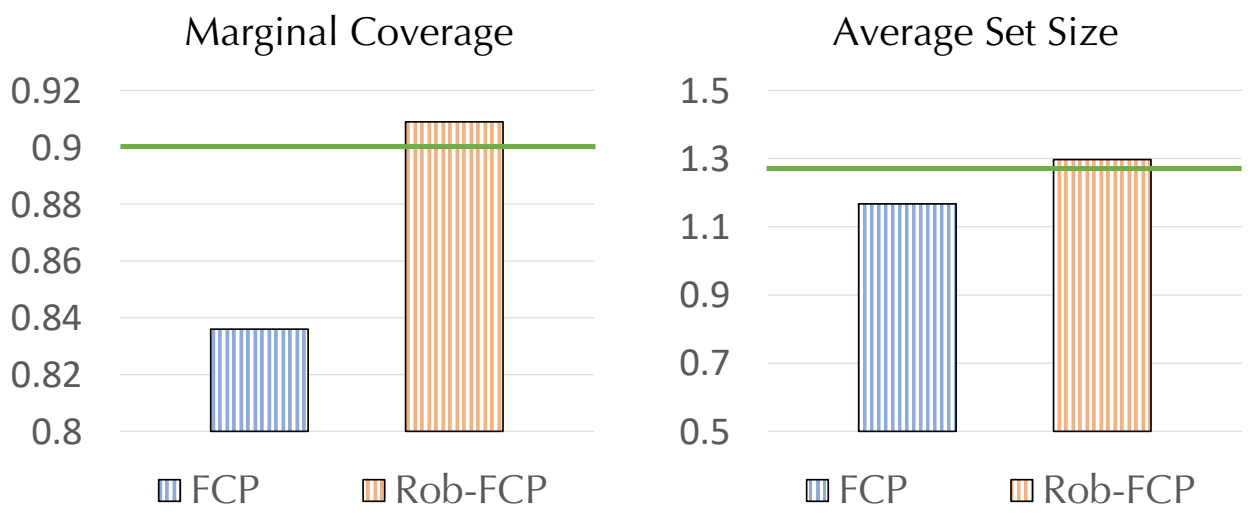


Figure 6: Marginal coverage / average set size under coverage attack with 40% malicious clients with $\beta = 0.0$ on CIFAR-10. The green horizontal line represents the benign marginal coverage and average set size without any malicious clients.

Ablation study of the selection of histogram granularity H We also add empirical evaluations to validate the trade-off of the selection of H in Table 8. The results in Table 8 demonstrate the empirical trade-off of the selection of dimensionality H and show that Rob-FCP remains effective for a broad range of H ($H = 10$ to $H = 1000$).

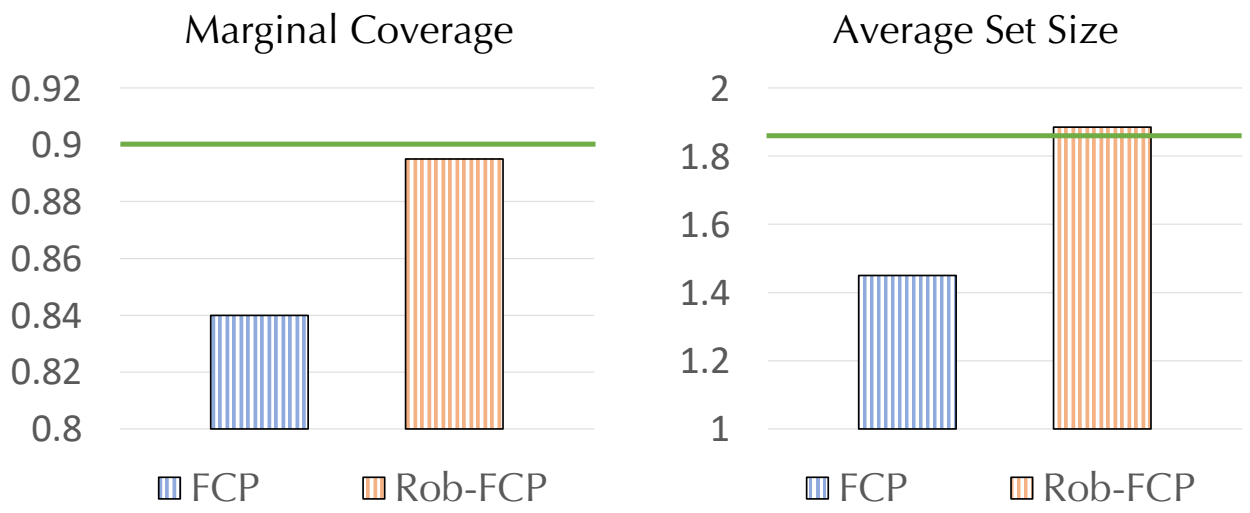


Figure 7: Marginal coverage / average set size under coverage attack with 40% malicious clients with $\beta = 0.5$ on CIFAR-10. The green horizontal line represents the benign marginal coverage and average set size without any malicious clients.

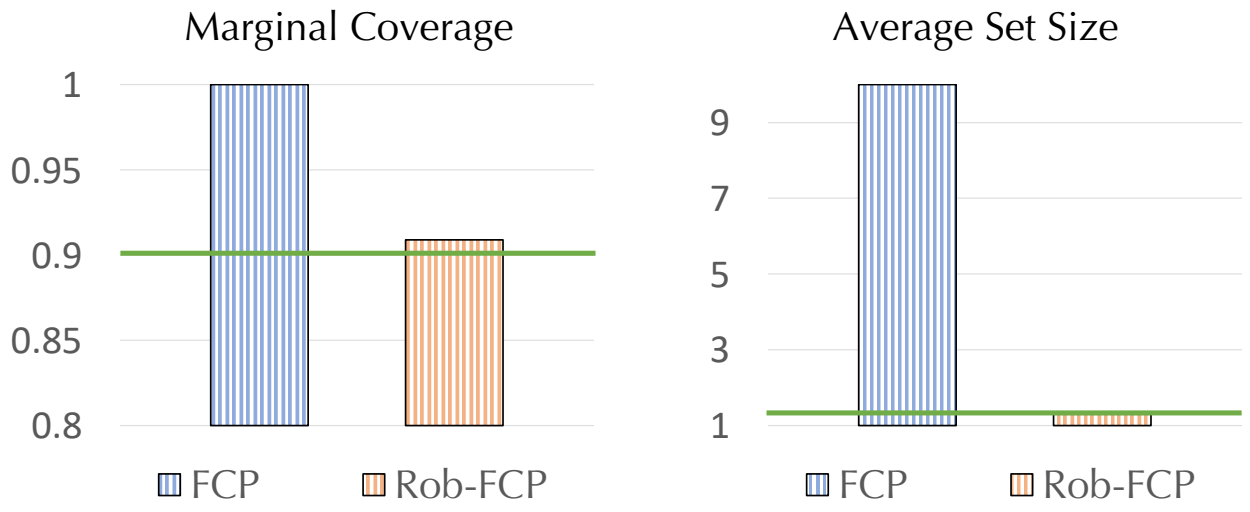


Figure 8: Marginal coverage / average set size under efficiency attack with 40% malicious clients with $\beta = 0.0$ on CIFAR-10. The green horizontal line represents the benign marginal coverage and average set size without any malicious clients.

Table 9: Marginal coverage / average set size under different Byzantine attacks with 10%, 20%, and 30% malicious clients. Rob-FCP consistently recovers the coverage (and average size of prediction set) of benign conformal prediction (Table 3), while the performance of FCP generally deteriorates as the percentage of malicious clients increases. β denotes the Dirichlet parameter for the partition of client data.

	Attack Method	Coverage Attack FCP	Coverage Attack Rob-FCP	Efficiency Attack FCP	Efficiency Attack Rob-FCP	Gaussian Attack FCP	Gaussian Attack Rob-FCP
$K_m/K = 10\%$							
$\beta = 0.0$	MNIST	0.896 / 0.898	**0.899 / 0.900**	0.999 / 4.034	**0.904 / 0.909**	0.947 / 0.960	**0.905 / 0.910**
	CIFAR-10	0.887 / 1.499	**0.900 / 1.588**	1.000 / 7.991	**0.892 / 1.556**	0.906 / 1.633	**0.892 / 1.565**
	T-ImageNet	0.873 / 18.44	**0.901 / 22.36**	0.999 / 148.7	**0.895 / 21.28**	0.916 / 23.98	**0.909 / 23.80**
	SHHS	0.889 / 1.303	**0.900 / 1.359**	0.999 / 5.338	**0.900 / 1.359**	0.909 / 1.409	**0.900 / 1.360**
	PathMNIST	0.892 / 1.184	**0.905 / 1.249**	1.000 / 6.271	**0.902 / 1.235**	0.941 / 1.504	**0.903 / 1.240**
$\beta = 0.5$	MNIST	0.892 / 1.747	**0.897 / 1.813**	1.000 / 9.319	**0.896 / 1.813**	0.892 / 1.798	**0.902 / 1.794**
	CIFAR-10	0.887 / 1.209	**0.894 / 2.287**	1.000 / 8.808	**0.908 / 2.347**	0.918 / 2.515	**0.911 / 2.378**
	T-ImageNet	0.892 / 41.03	**0.905 / 44.81**	0.997 / 146.7	**0.902 / 44.29**	0.917 / 47.47	**0.900 / 44.74**
	SHHS	0.889 / 1.304	**0.900 / 1.358**	1.000 / 5.981	**0.900 / 1.359**	0.909 / 1.412	**0.901 / 1.361**
	PathMNIST	0.892 / 1.290	**0.902 / 1.361**	0.996 / 5.149	**0.900 / 1.348**	0.938 / 1.739	**0.904 / 1.374**
$K_m/K = 20\%$							
$\beta = 0.0$	MNIST	0.873 / 0.876	**0.893 / 0.897**	1.000 / 10.00	**0.895 / 0.899**	0.967 / 0.988	**0.900 / 0.905**
	CIFAR-10	0.869 / 1.398	**0.888 / 1.532**	1.000 / 10.00	**0.913 / 1.659**	0.916 / 1.725	**0.903 / 1.633**
	T-ImageNet	0.874 / 17.787	**0.900 / 22.23**	1.000 / 200.0	**0.903 / 22.50**	0.908 / 23.12	**0.904 / 22.94**
	SHHS	0.876 / 1.243	**0.900 / 1.359**	1.000 / 5.984	**0.900 / 1.356**	0.918 / 1.467	**0.900 / 1.360**
	PathMNIST	0.880 / 1.134	**0.905 / 1.251**	1.000 / 8.335	**0.904 / 1.244**	0.983 / 2.434	**0.903 / 1.236**
$\beta = 0.5$	MNIST	0.857 / 1.534	**0.896 / 1.765**	1.000 / 9.089	**0.902 / 1.836**	0.915 / 1.945	**0.912 / 1.904**
	CIFAR-10	0.866 / 2.038	**0.896 / 2.314**	1.000 / 10.00	**0.908 / 2.366**	0.938 / 2.895	**0.892 / 2.256**
	T-ImageNet	0.860 / 33.99	**0.902 / 44.69**	1.000 / 199.0	**0.904 / 44.72**	0.922 / 49.44	**0.912 / 48.27**
	SHHS	0.874 / 1.236	**0.901 / 1.363**	1.000 / 5.985	**0.901 / 1.363**	0.917 / 1.463	**0.900 / 1.358**
	PathMNIST	0.876 / 1.210	**0.901 / 1.355**	1.000 / 7.395	**0.902 / 1.366**	0.980 / 2.905	**0.900 / 1.348**
$K_m/K = 30\%$							
$\beta = 0.0$	MNIST	0.851 / 0.854	**0.908 / 0.914**	1.000 / 10.00	**0.911 / 0.917**	0.977 / 1.009	**0.900 / 0.905**
	CIFAR-10	0.852 / 1.307	**0.895 / 1.583**	1.000 / 10.00	**0.894 / 1.563**	0.909 / 1.672	**0.903 / 1.602**
	T-ImageNet	0.862 / 15.66	**0.904 / 22.61**	1.000 / 200.0	**0.907 / 22.85**	0.907 / 23.89	**0.906 / 24.15**
	SHHS	0.859 / 1.176	**0.901 / 1.364**	1.000 / 6.000	**0.900 / 1.356**	0.926 / 1.526	**0.900 / 1.359**
	PathMNIST	0.863 / 1.064	**0.906 / 1.252**	1.000 / 9.000	**0.903 / 1.241**	1.000 / 6.531	**0.906 / 1.255**
$\beta = 0.5$	MNIST	0.849 / 1.451	**0.913 / 1.890**	1.000 / 10.00	**0.875 / 1.650**	0.925 / 2.010	**0.919 / 1.958**
	CIFAR-10	0.844 / 1.870	**0.900 / 2.294**	1.000 / 10.00	**0.912 / 2.408**	0.950 / 3.152	**0.901 / 2.327**
	T-ImageNet	0.864 / 33.41	**0.895 / 43.12**	1.000 / 200.0	**0.906 / 43.46**	**0.923 / 52.23**	0.932 / 55.78
	SHHS	0.857 / 1.169	**0.900 / 1.358**	1.000 / 6.000	**0.900 / 1.358**	0.927 / 1.530	**0.898 / 1.350**
	PathMNIST	0.860 / 1.141	**0.900 / 1.344**	1.000 / 9.000	**0.903 / 1.368**	1.000 / 6.287	**0.903 / 1.373**

Table 10: Marginal coverage / average set size under different Byzantine attacks with 10%, 20%, 30% and 40% malicious clients with unknown numbers of malicious clients. Rob-FCP consistently recovers the coverage (and average size of prediction set) of benign conformal prediction (Table 3), while the performance of FCP generally deteriorates as the percentage of malicious clients increases. β denotes the Dirichlet parameter for the partition of client data.

	Attack	Coverage Attack		Efficiency Attack		Gaussian Attack	
	Method	FCP	Rob-FCP	FCP	Rob-FCP	FCP	Rob-FCP
$K_m/K = 10\%$							
$\beta = 0.0$	MNIST	0.896 / 0.898	**0.901 / 0.905**	0.999 / 4.034	**0.890 / 0.895**	0.947 / 0.960	**0.895 / 0.900**
	CIFAR-10	0.887 / 1.499	**0.903 / 1.612**	1.000 / 7.991	**0.920 / 1.689**	0.906 / 1.633	**0.890 / 1.543**
	T-ImageNet	0.873 / 18.44	**0.908 / 22.52**	0.999 / 148.7	**0.890 / 20.93**	0.916 / 23.98	**0.897 / 21.64**
	SHHS	0.889 / 1.303	**0.902 / 1.365**	0.999 / 5.338	**0.903 / 1.368**	0.909 / 1.409	**0.902 / 1.367**
	PathMNIST	0.892 / 1.184	**0.899 / 1.237**	1.000 / 6.271	**0.905 / 1.253**	0.905 / 1.253	**0.901 / 1.239**
$\beta = 0.5$	MNIST	0.892 / 1.747	**0.895 / 1.798**	1.000 / 9.319	**0.900 / 1.780**	0.892 / 1.798	**0.896 / 1.800**
	CIFAR-10	0.887 / 1.209	**0.890 / 2.221**	1.000 / 8.808	**0.900 / 2.304**	0.918 / 2.515	**0.905 / 2.418**
	T-ImageNet	0.892 / 41.03	**0.903 / 43.94**	0.997 / 146.7	**0.898 / 43.01**	0.917 / 47.47	**0.915 / 47.35**
	SHHS	0.889 / 1.304	**0.902 / 1.367**	1.000 / 5.981	**0.902 / 1.364**	0.909 / 1.412	**0.900 / 1.357**
	PathMNIST	**0.892** / 1.290	0.909 / **1.394**	0.996 / 5.149	**0.901 / 1.376**	**0.905** / 1.387	0.907 / **1.375**
$K_m/K = 20\%$							
$\beta = 0.0$	MNIST	0.873 / 0.876	**0.898 / 0.903**	1.000 / 10.00	**0.906 / 0.912**	0.967 / 0.988	**0.904 / 0.908**
	CIFAR-10	0.869 / 1.398	**0.888 / 1.512**	1.000 / 10.00	**0.902 / 1.603**	0.916 / 1.725	**0.905 / 1.623**
	T-ImageNet	0.874 / 17.787	**0.904 / 22.47**	1.000 / 200.0	**0.907 / 22.76**	0.908 / 23.12	**0.904 / 22.88**
	SHHS	0.876 / 1.243	**0.902 / 1.365**	1.000 / 5.984	**0.902 / 1.366**	0.918 / 1.467	**0.902 / 1.363**
	PathMNIST	0.880 / 1.134	**0.900 / 1.229**	1.000 / 8.335	**0.902 / 1.241**	0.909 / 1.273	**0.898 / 1.229**
$\beta = 0.5$	MNIST	0.857 / 1.534	**0.901 / 1.832**	1.000 / 9.089	**0.881 / 1.713**	0.915 / 1.945	**0.908 / 1.889**
	CIFAR-10	0.866 / 2.038	**0.900 / 2.344**	1.000 / 10.00	**0.897 / 2.312**	0.938 / 2.895	**0.929 / 2.702**
	T-ImageNet	0.860 / 33.99	**0.905 / 44.38**	1.000 / 199.0	**0.894 / 42.30**	0.922 / 49.44	**0.906 / 46.38**
	SHHS	0.874 / 1.236	**0.901 / 1.362**	1.000 / 5.985	**0.903 / 1.369**	0.917 / 1.463	**0.902 / 1.365**
	PathMNIST	0.876 / 1.210	**0.907 / 1.388**	1.000 / 7.395	**0.903 / 1.362**	0.905 / 1.382	**0.902 / 1.362**
$K_m/K = 30\%$							
$\beta = 0.0$	MNIST	0.851 / 0.854	**0.905 / 0.912**	1.000 / 10.00	**0.907 / 0.913**	0.977 / 1.009	**0.903 / 0.908**
	CIFAR-10	0.852 / 1.307	**0.904 / 1.612**	1.000 / 10.00	**0.891 / 1.544**	0.909 / 1.672	**0.903 / 1.578**
	T-ImageNet	0.862 / 15.66	**0.902 / 21.92**	1.000 / 200.0	**0.903 / 22.19**	0.907 / 23.89	**0.906 / 23.77**
	SHHS	0.859 / 1.176	**0.903 / 1.372**	1.000 / 6.000	**0.902 / 1.366**	0.926 / 1.526	**0.903 / 1.368**
	PathMNIST	0.863 / 1.064	**0.902 / 1.239**	1.000 / 9.000	**0.898 / 1.221**	0.907 / 1.263	**0.905 / 1.246**
$\beta = 0.5$	MNIST	0.849 / 1.451	**0.920 / 1.947**	1.000 / 10.00	**0.900 / 1.779**	0.925 / 2.010	**0.911 / 1.943**
	CIFAR-10	0.844 / 1.870	**0.899 / 2.360**	1.000 / 10.00	**0.891 / 2.264**	0.950 / 3.152	**0.896 / 2.300**
	T-ImageNet	0.864 / 33.41	**0.895 / 42.79**	1.000 / 200.0	**0.908 / 44.74**	0.923 / 52.23	**0.920 / 50.70**
	SHHS	0.857 / 1.169	**0.902 / 1.368**	1.000 / 6.000	**0.904 / 1.374**	0.927 / 1.530	**0.903 / 1.370**
	PathMNIST	0.860 / 1.141	**0.895 / 1.337**	1.000 / 9.000	**0.902 / 1.376**	0.910 / 1.418	**0.899 / 1.352**
$K_m/K = 40\%$							
$\beta = 0.0$	MNIST	0.832 / 0.834	**0.891 / 0.892**	1.000 / 10.00	**0.895 / 0.901**	0.979 / 1.025	**0.899 / 0.904**
	CIFAR-10	0.831 / 1.189	**0.913 / 1.666**	1.000 / 10.00	**0.902 / 1.608**	0.916 / 1.733	**0.905 / 1.612**
	T-ImageNet	0.830 / 12.97	**0.888 / 21.45**	1.000 / 200.0	**0.905 / 22.99**	0.918 / 25.69	**0.903 / 23.42**
	SHHS	0.834 / 1.093	**0.902 / 1.363**	1.000 / 6.000	**0.903 / 1.369**	0.937 / 1.611	**0.902 / 1.368**
	PathMNIST	0.840 / 0.997	**0.901 / 1.246**	1.000 / 9.000	**0.898 / 1.237**	0.914 / 1.302	**0.899 / 1.250**
$\beta = 0.5$	MNIST	0.805 / 1.284	**0.911 / 1.929**	1.000 / 10.00	**0.910 / 1.906**	0.941 / 2.227	**0.929 / 2.084**
	CIFAR-10	0.829 / 1.758	**0.893 / 2.270**	1.000 / 10.00	**0.888 / 2.203**	0.970 / 3.863	**0.923 / 2.635**
	T-ImageNet	0.825 / 27.84	**0.906 / 45.18**	1.000 / 200.0	**0.903 / 42.62**	0.942 / 61.50	**0.937 / 59.61**
	SHHS	0.835 / 1.095	**0.902 / 1.364**	1.000 / 6.000	**0.904 / 1.375**	0.937 / 1.609	**0.903 / 1.371**
	PathMNIST	0.837 / 1.055	**0.903 / 1.378**	1.000 / 9.000	**0.909 / 1.398**	0.915 / 1.464	**0.914 / 1.488**

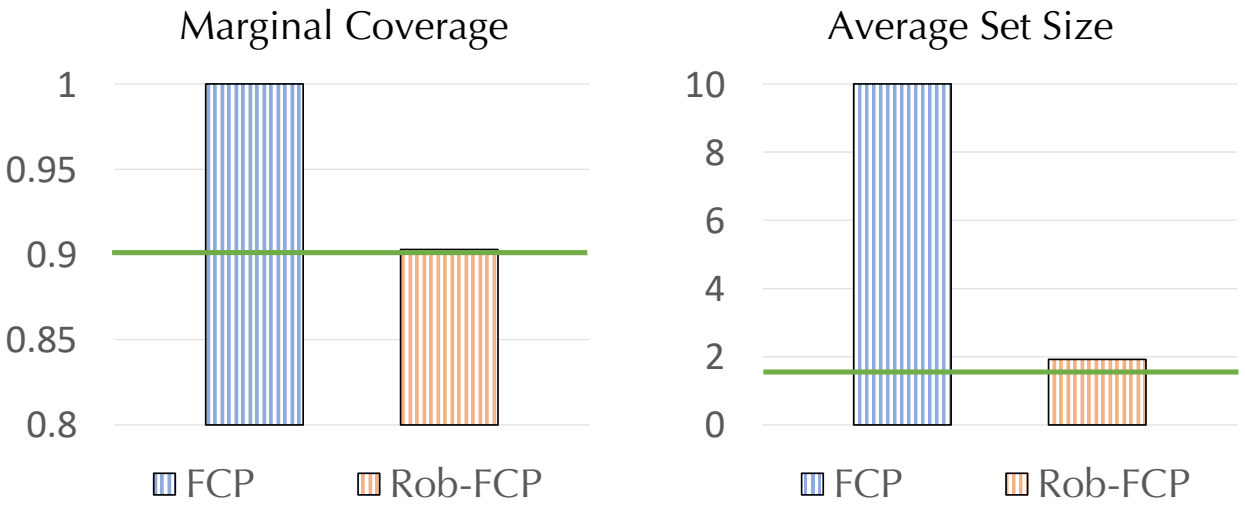


Figure 9: Marginal coverage / average set size under efficiency attack with 40% malicious clients with $\beta = 0.5$ on CIFAR-10. The green horizontal line represents the benign marginal coverage and average set size without any malicious clients.

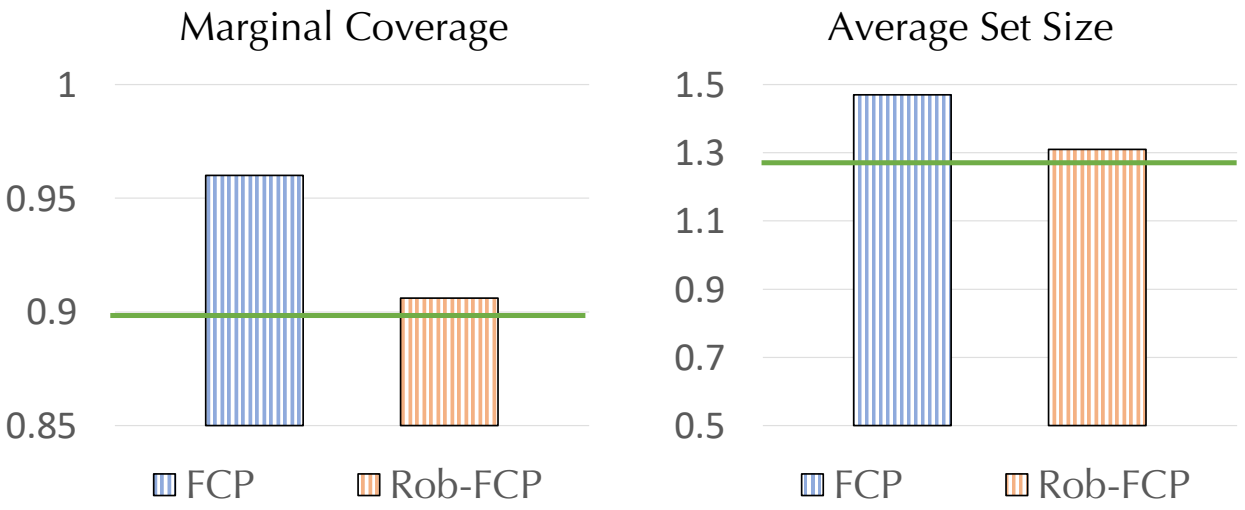


Figure 10: Marginal coverage / average set size under Gaussian attack with 40% malicious clients with $\beta = 0.0$ on CIFAR-10. The green horizontal line represents the benign marginal coverage and average set size without any malicious clients.

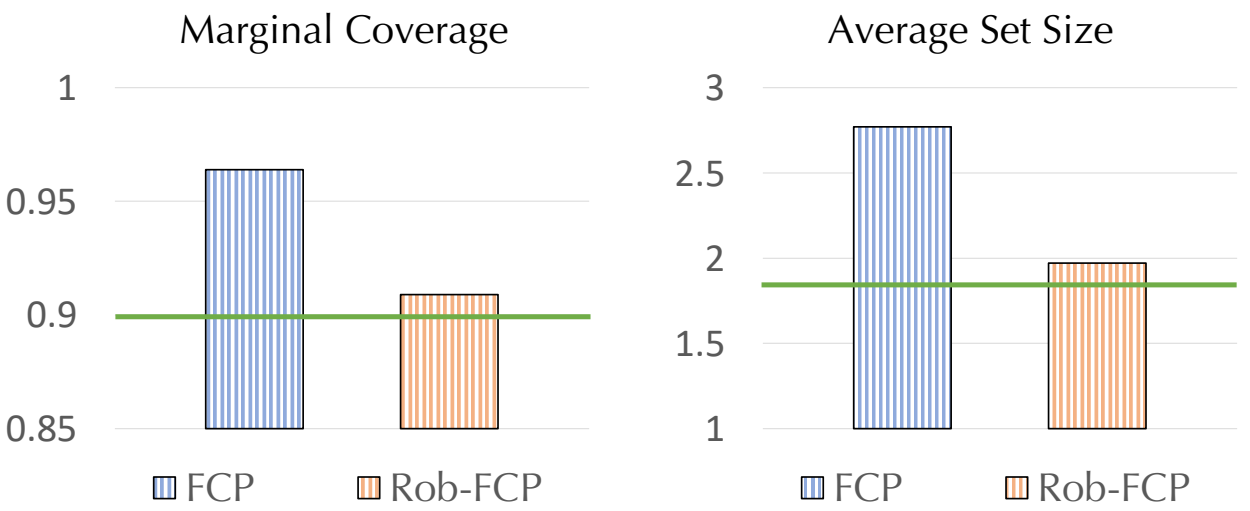


Figure 11: Marginal coverage / average set size under Gaussian attack with 40% malicious clients with $\beta = 0.5$ on CIFAR-10. The green horizontal line represents the benign marginal coverage and average set size without any malicious clients.

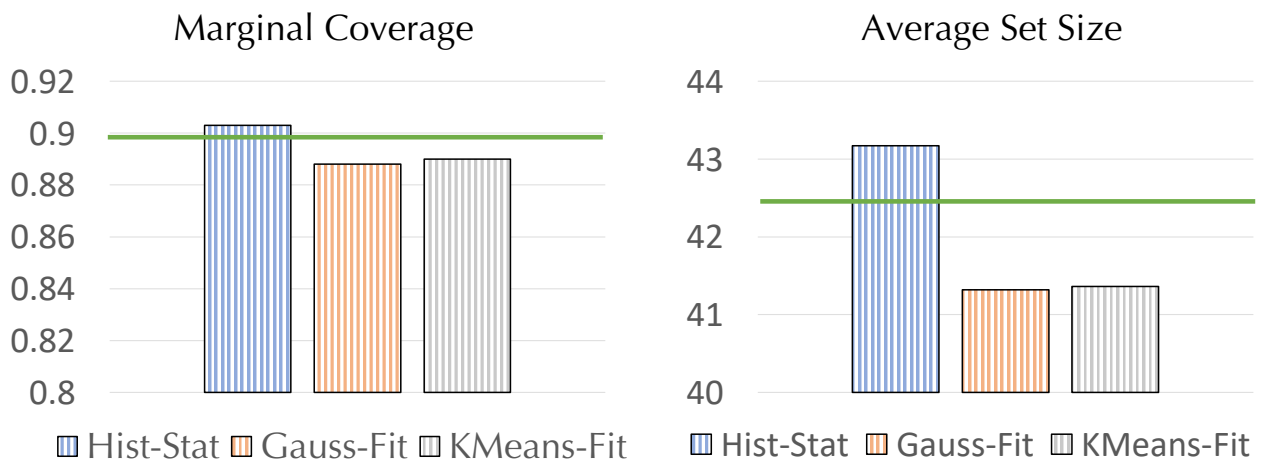


Figure 12: Marginal coverage / average set size under coverage attack with 40% malicious clients with $\beta = 0.5$ on Tiny-ImageNet. The green horizontal line represents the benign marginal coverage and average set size without any malicious clients.

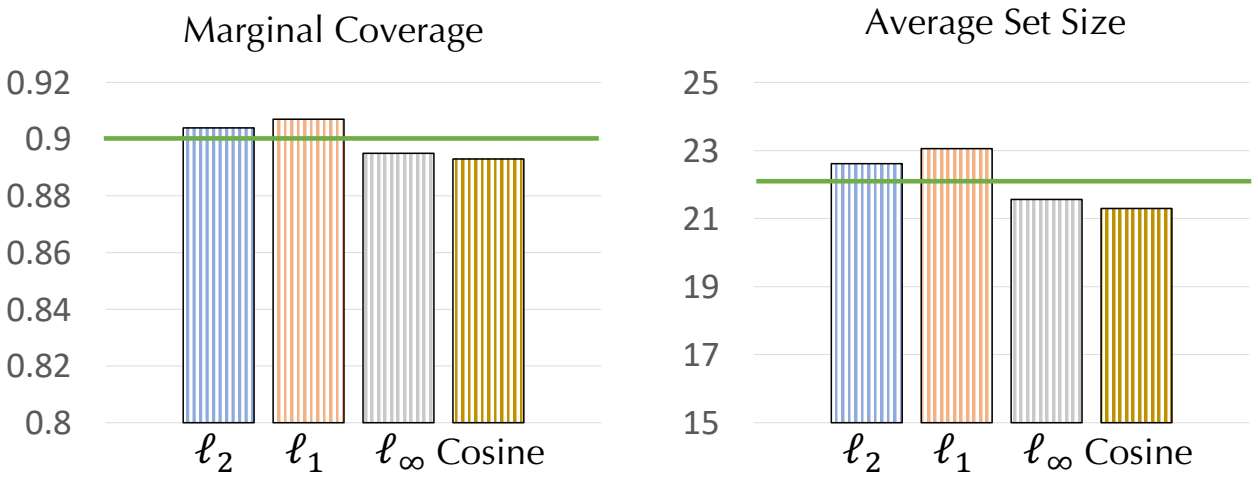


Figure 13: Marginal coverage / average set size under coverage attack with 40% malicious clients with $\beta = 0.0$ on Tiny-ImageNet. The green horizontal line represents the benign marginal coverage and average set size without any malicious clients.