Appendix

A Neural Network configurations

We train these models with a batch size of 10, an SGD optimizer operates with a learning rate of 0.01, and 0.5 momentum for 25 epochs. The accuracy of the model is evaluated on a holdout set of 1000 samples.

A.1 Fashion-MNIST

- Layer 1: $1 * 16 * 5$, 2D Convolution, Batch Normalization, ReLU Activation, Max pooling.
- Layer 2: $16*32*5$, 2D Convolution, Batch Normalization, ReLU Activation, Max pooling.
- Output: 10 Classes, Linear.

A.2 Cifar-10

- Layer 1: $1 * 32 * 3$, 2D Convolution, Batch Normalization, ReLU Activation, Max pooling.
- Layer 2: $32*32*3$, 2D Convolution, Batch Normalization, ReLU Activation, Max pooling.
- Output: 10 Classes, Linear.

B Metrics

The metrics observed in Section 3 to evaluate the performance of the defense mechanisms are defined as follows:

$$\begin{aligned}
&\text{Precision} = \frac{\text{TP}}{\text{TP+FP}},\\
&\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}},\\
&\text{Recall} = \frac{\text{TP}}{\text{TP+FN}},\\
&\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}.
\end{aligned}$$

C Byanzine Attack Definitions

C.1 Gaussian Attack (GA)

Definition 1 (Gaussian attack)**.** *In a Gaussian attack, the attacker manipulates malicious nodes to send Gaussian random messages to the global coordinator, i.e., $\{\boldsymbol{M}_{i,:}\}_{i\in\mathcal{I}_m}$ are independent random samples from Gaussian distribution $\mathcal{MVN}(\boldsymbol{m}_b, \Sigma)$, where $\boldsymbol{m}_b = \frac{1}{n_1}\sum_{i\in\mathcal{I}_b} \boldsymbol{M}_{i,:}$ and Σ is the covariance matrix determined by the attacker.*

C.2 Sign Flipping Attack (SF)

Definition 2 (Sign flipping attack)**.** *Sign flipping attack aims to generate the gradient values of malicious nodes by flipping the sign of the average of all the benign nodes' gradient at each epoch, i.e., specifying $\boldsymbol{M}_{i,:} = -r\boldsymbol{m}_b$ for any $i \in \mathcal{I}_m$, where $r > 0, \boldsymbol{m}_b = \frac{1}{n_1}\sum_{k\in\mathcal{I}_b} \boldsymbol{M}_{k,:}$.*

C.3 Zero Gradient Attack (ZG)

Definition 3 (Zero gradient attack)**.** *Zero gradient attack aims to make the aggregated message to be zero, i.e., $\sum_{i=1}^{n} \boldsymbol{M}_{i,:} = 0$, at each epoch, by specifying $\boldsymbol{M}_{i,:} = -\frac{n_1}{n_0}\boldsymbol{m}_b$ for all $i \in \mathcal{I}_m$.*

D Proof of Theorem 1

Proof. To prove Equation 5, we start with deriving the expectation of v_i for benign nodes. Let v_i^b denote the variance of a benign node.

$$\begin{aligned}\lim_{p\to\infty}\mathbb{E}(v_i^b) &= \lim_{p\to\infty}\mathbb{E}\left(\frac{1}{p}\sum_{j=1}^{p}(\boldsymbol{R}_{i,j}-e_i)^2\right)\\ &= \lim_{p\to\infty}\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}(\boldsymbol{R}_{i,j}-e_i)^2)\\ &= \frac{1}{p}\sum_{j=1}^{p}\mathrm{Var}(\boldsymbol{R}_{i,j}) = \bar{v}^b, \qquad (10)\end{aligned}$$

where $\bar{v}^b = \frac{1}{p}\sum_{j=1}^{p}\mathrm{Var}(\boldsymbol{R}_{i\in\mathcal{I}_b,j})$ is the average variance of a benign node across p dimensions.

The variance of v_i^b is given as,

$$\begin{aligned}\lim_{p\to\infty}\mathrm{Var}(v_i^b) &= \lim_{p\to\infty}\mathrm{Var}(\frac{1}{p}\sum_{j=1}^{p}(\boldsymbol{R}_{i,j}-e_i)^2)\\ &= \lim_{p\to\infty}\frac{1}{p^2}\sum_{j=1}^{p}\mathrm{Var}((\boldsymbol{R}_{i,j}-e_i)^2)\\ &= 0 \qquad (11)\end{aligned}$$

Equation 11 holds because $\mathrm{Var}((\boldsymbol{R}_{i,j}-e_i)^2) < \infty$. Combining Equation 10 and 11, we have

$$\lim_{p\to\infty} v_i^b = \lim_{p\to\infty}\mathbb{E}(v_i^b) = \bar{v}_b. \qquad (12)$$

It completes the proof for benign nodes. Considering that the proof for malicious nodes can be derived in exactly the same way, the proof is complete.

□

E Independence test

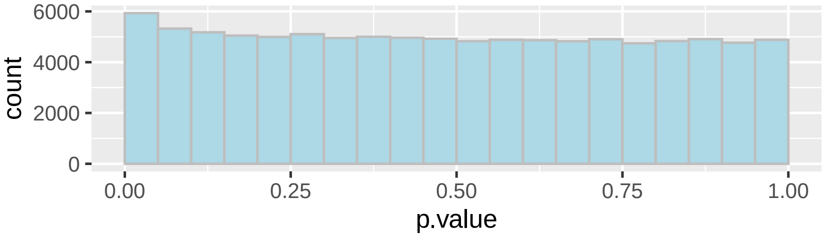


Figure 4: Independence tests for 100000 column pairs randomly chosen from the message matrix $\boldsymbol{M}$ generated from the MNIST data provides support to the independence assumption made in Theorem 1.

F Proof of Lemma 3

Proof. For a malicious node, according to the definition of sign flipping attack,

$$\boldsymbol{M}_{i,j} = -\frac{r}{n_1}\sum_{i\in\mathcal{I}_b}\boldsymbol{M}_{i,j}. \qquad (13)$$

As all the benign nodes are distributed as normal distributions. The malicious nodes also follow normal distributions. And the expectation and variance are

$$g_j = \mathbb{E}(\boldsymbol{M}_{i,j}) = -\frac{r}{n_1}\sum_{i\in\mathcal{I}_b}\mathbb{E}(\boldsymbol{M}_{i,j}) = -rf_j, \tag{14}$$

$$\sigma_{m,j}^2 = \mathrm{Var}(\boldsymbol{M}_{i,j}) = \frac{r^2}{n_1^2}\sum_{i\in\mathcal{I}_b}\mathrm{Var}(\boldsymbol{M}_{i,j}) = \frac{r^2}{n_1}\sigma_j^2. \tag{15}$$

Thus, it completes the proof. □

G Proof of Theorem 2

Proof. Given the fact that $\lim_{N\to\infty}\sigma_j^2 = 0$, all the messages from benign nodes will converge to their expectation which is μ_j. Mathematically,

$$\lim_{N_{min}\to\infty}\boldsymbol{M}_{i,j} = \mu_j, i\in\mathcal{I}_b. \tag{16}$$

According to the definition of sign flipping attack, for a malicious node,

$$\lim_{N_{min}\to\infty}\boldsymbol{M}_{i,j} = -r\mu_j, i\in\mathcal{I}_m. \tag{17}$$

Combining Equation 16, 17 and the definition of *Rank* operator, we have

$$\lim_{N_{min}\to\infty}\mathbb{E}(\boldsymbol{R}_{i,j}) = \left\{\begin{array}{cc} \frac{1+n_0}{2}, i\in\mathcal{I}_m, & \text{if } \mu_j < 0, \\ \frac{1+n_1+n}{2}, i\in\mathcal{I}_m, & \text{if } \mu_j > 0, \end{array}\right.$$

and

$$\lim_{N_{min}\to\infty}\boldsymbol{R}_{i,j} \sim \left\{\begin{array}{cc} \mathrm{Unif}(n_0+1, n), i\in\mathcal{I}_b, & \text{if } \mu_j < 0, \\ \mathrm{Unif}(1, n_1), i\in\mathcal{I}_b, & \text{if } \mu_j > 0, \end{array}\right.$$

leading to

$$\lim_{N_{min}\to\infty}\mathbb{E}(\boldsymbol{R}_{i,j}) = \left\{\begin{array}{cc} \frac{n_0+1+n}{2}, i\in\mathcal{I}_b, & \text{if } \mu_j < 0, \\ \frac{1+n_1}{2}, i\in\mathcal{I}_b, & \text{if } \mu_j > 0, \end{array}\right.$$

where Unif(a, b) denotes the uniform distribution over the integer space in $[a, b]$.

Then, according to the strong law of large numbers, we have,

$$\lim_{N_{min}\to\infty} e_i = \left\{\begin{array}{cc} \frac{1}{p}(p_+(\frac{n+n_1+1}{2}) + (p-p_+)\frac{n_0+1}{2}) = n_1\frac{p_+}{p} + \frac{n_0+1}{2},\ a.s. & i\in\mathcal{I}_m, \\ \frac{1}{p}(p_+(\frac{n_1+1}{2}) + (p-p_+)\frac{n+n_0+1}{2}) = \frac{n+n_0+1}{2} - n_0\frac{p_+}{p},\ a.s. & i\in\mathcal{I}_b, p\to\infty. \end{array}\right. \tag{18}$$

Define $\bar{a}_b \triangleq \frac{n+n_0+1}{2} - n_0\frac{p_+}{p}$ and $\bar{a}_m \triangleq n_1\frac{p_+}{p} + \frac{n_0+1}{2}$. It completes the proof of Equation 8.

To prove Equation 9, we start with computing the expectation of the variance of a malicious node $i\in\mathcal{I}_m$.

$$\begin{aligned} & \mathbb{E}(\lim_{N_{min}\to\infty} v_i) = \\ = \ & \frac{1}{p}\mathbb{E}\left(\lim_{N_{min}\to\infty}\sum_{j=1}^{p}(\boldsymbol{R}_{i,j} - e_i)^2\right) \\ = \ & \frac{1}{p}\left(p_+(\frac{n+n_1+1}{2} - e_i)^2 + (p-p_+)(\frac{1+n_0}{2} - e_i)^2\right) \\ = \ & \frac{1}{p}\left(p_+(\frac{n+n_1+1}{2} - (n_1\frac{p_+}{p} + \frac{n_0+1}{2}))^2 + (p-p_+)(\frac{1+n_0}{2} - (n_1\frac{p_+}{p} + \frac{n_0+1}{2}))^2\right) \\ = \ & \frac{p_+(p-p_+)n_1^2}{p^2} = \bar{v}_m \end{aligned}$$

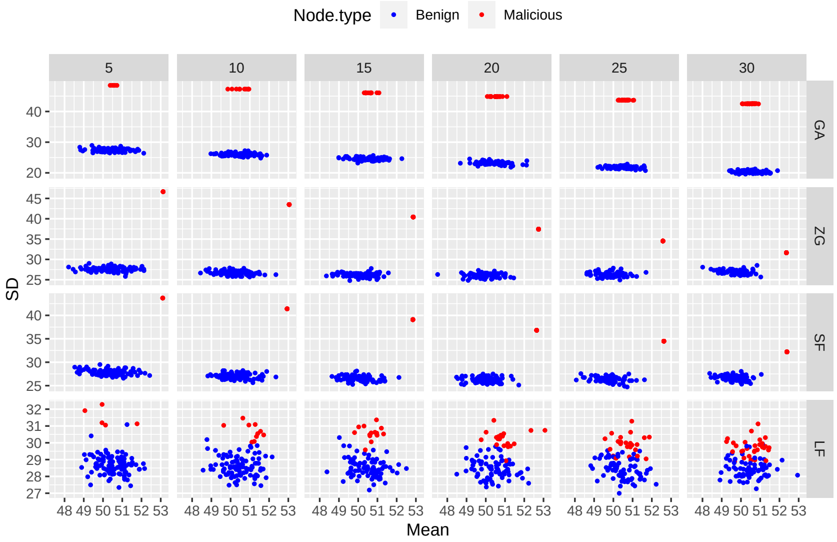


Figure 5: The scatter plots of (e_i, s_i) for the 100 nodes under four types of attack as illustrative examples demonstrating ranking mean and variance from the 1st epoch of training for the FASHION-MNIST dataset.

Similarly, the expectation of the variance of a benign node is computed as,

$$\begin{aligned}
&\mathbb{E}(\lim_{p\to\infty}\lim_{N_{min}\to\infty} v_i)\\
=\ &\frac{1}{p}\mathbb{E}\left(\lim_{p\to\infty}\lim_{N_{min}\to\infty}\sum_{j=1}^{p}(\boldsymbol{R}_{i,j}-e_i)^2\right)\\
=\ &\frac{1}{p}\left(p_+\frac{1}{n_1}\sum_{k=1}^{n_1}(k-e_i)^2+(p-p_+)\frac{1}{n_1}\sum_{k=n_0+1}^{n}(k-e_i)^2\right)\\
=\ &\beta_0+\beta_1 p_+ +\beta_2 p_+^2=\bar{v}_b,
\end{aligned}$$

where $\beta_0=\frac{7n^2+7n_0^2+10nn_0+2n+2n_0-1}{12}$, $\beta_1=\frac{n_0}{p}(2n+3n_0+2)$ and $\beta_2=-\frac{n_0^2}{p^2}$.

According to the strong law of large numbers, we have Equation 9. It completes the proof. □

H Proof of Corollary 1

In both of the zero gradient attack and sign flipping attack, the malicious nodes generate their gradient values based on the average of the benign nodes' gradient. The only difference is what magnitude the malicious nodes adopt. And zero gradient attack is the special case of sign flipping attack with $u=\frac{n_1}{n_0}$. However, when the message matrix is transformed to the permutation space, the impact of the value of u disappears. Therefore, the behaviour of all the nodes in the permutation space does not change as the attack type changes from sign flipping to zero gradient. The proof of Corollary 1 is same as Theorem 2.

I Additional experimentation

I.1 Illustration of the average ranking and variance of ranking

Section 2 speculated that the distribution of parameter ranks differ sufficiently for the detection of malicious and benign nodes. We validate this hypothesis in figure 5 by illustrating the difference between the benign nodes and malicious nodes in terms of the mean of gradients' rankings and the variance of gradients' ranking.

It can be observed from Figure 5 that, under GA and LF attacks, the average rankings of malicious nodes are of a similar distribution to benign nodes. It is problematic for distinguishing between the two types of nodes, if only average ranking information is used. On the other hand, Figure 5 displays a larger separation of distributions for the variance of ranking. It is noted that all 4 attacks observe a convergence of the distributions as the number of malicious nodes increase, increasing the difficulty of defense for both MANDERA and all other defenses. However, the likelihood of an attacker controlling increasingly large numbers of malicious nodes also decrease.

I.2 Model Loss

Figure 6 presents the model loss to accompany the model prediction performance of Figure 3 previously seen in Section 3.

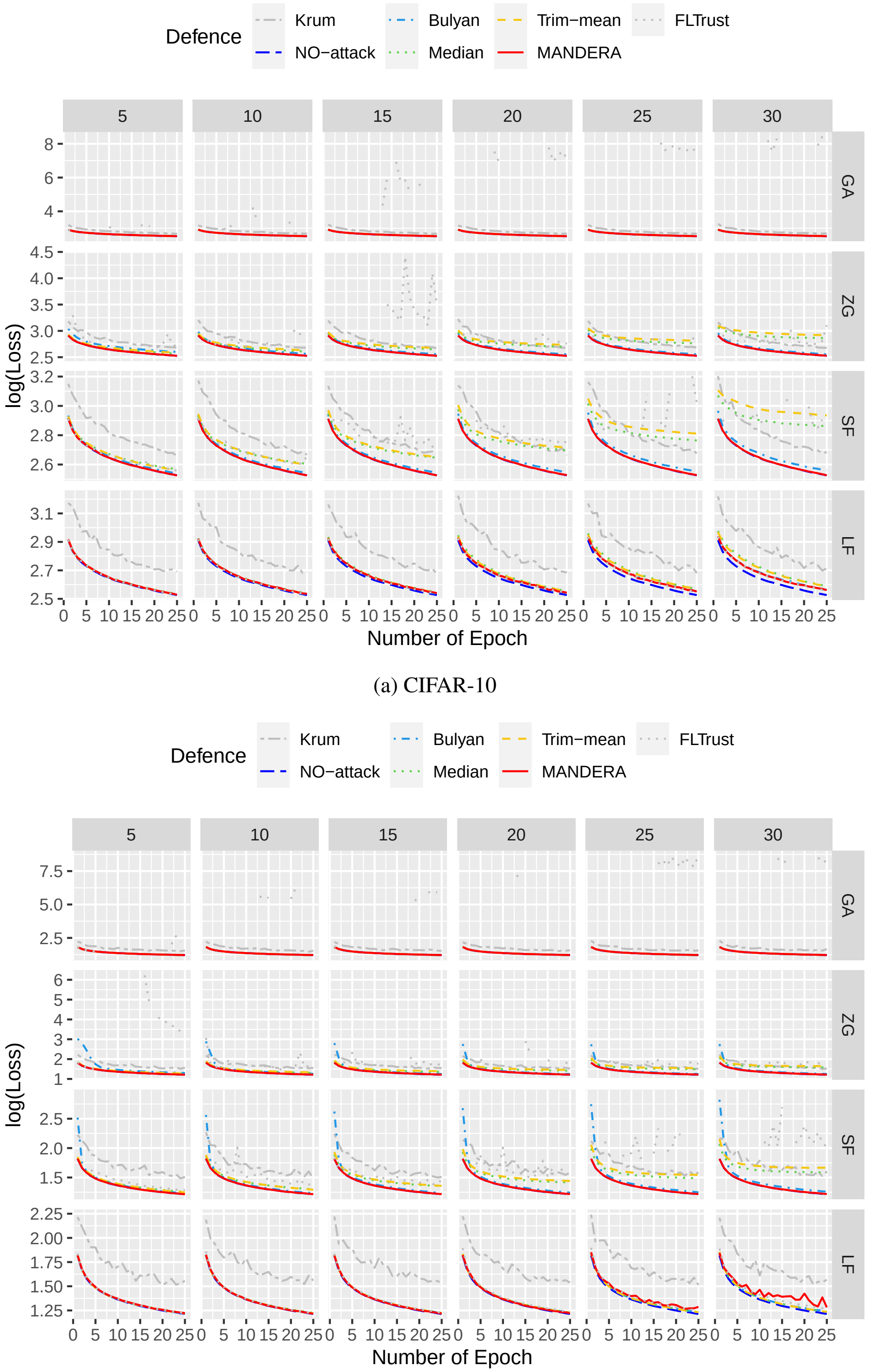
I.3 Computational Efficiency

We have previously been able to observe that MANDERA can perform at par with the current highest performing poisoning attack defenses. Another benefit arises with the simplification of the mitigation strategy with the introduction of ranking at the core of the algorithm. Sorting and Ranking algorithms are fast. Additionally, we only apply clustering on the two dimensions of rank mean and standard deviation, in contrast to other works that seek to cluster on the entire node update [Chen et al., 2021]. The timings of Table 1 for MANDERA, Krum and Bulyan do not include the parameter/gradient aggregation step. These timings were computed on 1 core of a Dual Xeon 14-core E5-2690, with 8 gb of system RAM and a single NVidia Tesla P100. Table 1 demonstrates that MANDERA is able to achieve a faster speed than that of single Krum [2] (by more than half) and Bulyan (by an order of magnitude).

Table 1: Mean and standard deviation of defense function timings given the same set of gradients from 100 nodes, of which 30 were malicious. Each function was repeated 100 times.

Defense (Detection)	Mean ± Std (ms)	Defense (Aggregation)	Mean ± Std (ms)
MANDERA	643 ± 8.646	Trimmed Mean	3.96 ± 0.41
Krum (Single)	1352 ± 10.09	Median	9.81 ± 3.88
Bulyan	27209 ± 233.4		

[2] The use of multi-krum would have yielded better protection (c.f. Section 3) at the behest of speed.



(a) CIFAR-10

(b) FASHION-MNIST

Figure 6: Model Loss at each epoch of training, each line of the curve represents a different defense against the attacks (GA: Gaussian attack; ZG: Zero-gradient attack; SF: Sign-flipping; and LF: Label-flipping).