A. Convergence Analysis of Angular Trimmed-mean Aggregation (ATM)

Before proving Theorem 1, we first present Lemma 1. The proof is partially inspired by [49].

Lemma 1. *Let $\{\theta_i\}_{i=1}^n$ be a sorted sequence of scalar values in ascending order, where m entries are assumed to be malicious. For clarity, we refer to the remaining $n-m$ benign values as $\{\hat{\theta}_i\}_{i=1}^{n-m}$, which form a subset of the original sequence. Thus, for $m < b \leq \lfloor n/2 \rfloor - 1$,*

$$\hat{\theta}_{b-m+i} \overset{(I)}{\leq} \theta_{b+i} \overset{(II)}{\leq} \hat{\theta}_{b+i}, \quad 1 \leq i \leq n-2b,$$

where $\hat{\theta}_{b+i}$ is the $(b+i)$-th smallest element in $\{\hat{\theta}_i\}_{i=1}^{n-m}$, and θ_{b+i} is the $(b+i)$-th smallest element in $\{\theta_i\}_{i=1}^n$.

Proof. We prove each of the two inequalities individually.

Inequality I: Suppose for contradiction that $\hat{\theta}_{b-m+i} > \theta_{b+i}$. This implies there exist $(n-m)-(b-m+i)+1 = n-b-i+1$ correct values strictly greater than θ_{b+i}. However, since θ_{b+i} is the $(b+i)$-th smallest element in $\{\theta_i\}_{i=1}^n$, there can be at most $n-(b+i) = n-b-i$ elements greater than θ_{b+i}. This contradiction establishes $\hat{\theta}_{b-m+i} \leq \theta_{b+i}$.

Inequality II: Suppose for contradiction that $\theta_{b+i} > \hat{\theta}_{b+i}$. This implies there exist $b+i$ correct values strictly less than θ_{b+i}. However, since θ_{b+i} is the $(b+i)$-th smallest element in$\{\theta_i\}_{i=1}^n$, there can be at most $b+i-1$ elements less than θ_{b+i}. This contradiction establishes $\theta_{b+i} \leq \hat{\theta}_{b+i}$. □

Proof of Theorem 1: According to Lemma 1, we have

$$\begin{aligned}
&\sum_{i=b-m+1}^{n-b-m}(\hat{\theta}_i-\omega) \leq \sum_{i=b+1}^{n-b}(\theta_i-\omega) \leq \sum_{i=b+1}^{n-b}(\hat{\theta}_i-\omega) \\
\Rightarrow\quad & \frac{\sum_{i=1}^{n-b-m}(\hat{\theta}_i-\omega)}{n-b-m} \leq \frac{\sum_{i=b+1}^{n-b}(\theta_i-\omega)}{n-2b} \leq \frac{\sum_{i=b+1}^{n-m}(\hat{\theta}_i-\omega)}{n-b-m} \\
\Rightarrow\quad & \left[\frac{\sum_{i=b+1}^{n-b}(\theta_i-\omega)}{n-2b}\right]^2 \\
& \leq \max\left\{\left[\frac{\sum_{i=1}^{n-b-m}(\hat{\theta}_i-\omega)}{n-b-m}\right]^2, \left[\frac{\sum_{i=b+1}^{n-m}(\hat{\theta}_i-\omega)}{n-b-m}\right]^2\right\}.
\end{aligned}$$

Thus, one has that:

$$\begin{aligned}
&\left[\frac{1}{|\mathcal{G}'|}\sum_{\boldsymbol{g}\in\mathcal{G}'}\theta_{\boldsymbol{g}}-\omega\right]^2 \\
&= \left[\frac{\sum_{i=b+1}^{n-b}\theta_i}{n-2b}-\omega\right]^2 \\
&= \left[\frac{\sum_{i=b+1}^{n-b}(\theta_i-\omega)}{n-2b}\right]^2 \\
&\leq \max\left\{\left[\frac{\sum_{i=1}^{n-b-m}(\hat{\theta}_i-\omega)}{n-b-m}\right]^2, \left[\frac{\sum_{i=b+1}^{n-m}(\hat{\theta}_i-\omega)}{n-b-m}\right]^2\right\}.
\end{aligned}$$

Note that for any subset $T \subseteq [n-m]$ with size $|T| = n-b-m$, the following bound holds:

$$\begin{aligned}
&\left[\frac{\sum_{i\in T}(\hat{\theta}_i-\omega)}{n-b-m}\right]^2 \\
&= \left[\frac{\sum_{i\in[n-m]}(\hat{\theta}_i-\omega)-\sum_{i\notin T}(\hat{\theta}_i-\omega)}{n-b-m}\right]^2 \\
&\leq 2\left[\frac{\sum_{i\in[n-m]}(\hat{\theta}_i-\omega)}{n-b-m}\right]^2 + 2\left[\frac{\sum_{i\notin T}(\hat{\theta}_i-\omega)}{n-b-m}\right]^2 \\
&= \frac{2(n-m)^2}{(n-b-m)^2}\left[\frac{\sum_{i\in[n-m]}(\hat{\theta}_i-\omega)}{n-m}\right]^2 \\
&\quad + \frac{2b^2}{(n-b-m)^2}\left[\frac{\sum_{i\notin T}(\hat{\theta}_i-\omega)}{b}\right]^2 \\
&\leq \frac{2(n-m)^2}{(n-b-m)^2}\left[\frac{\sum_{i\in[n-m]}(\hat{\theta}_i-\omega)}{n-m}\right]^2 \\
&\quad + \frac{2b^2}{(n-b-m)^2}\frac{\left[\sum_{i\notin T}(\hat{\theta}_i-\omega)\right]^2}{b} \\
&\leq \frac{2(n-m)^2}{(n-b-m)^2}\left[\frac{\sum_{i\in[n-m]}(\hat{\theta}_i-\omega)}{n-m}\right]^2 \\
&\quad + \frac{2b^2}{(n-b-m)^2}\frac{\left[\sum_{i\in[n-m]}(\hat{\theta}_i-\omega)\right]^2}{b}.
\end{aligned}$$

Taking the expectation yields:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\sum_{i\in T}(\hat{\theta}_i-\omega)}{n-b-m}\right]^2 \\
&\leq \frac{2(n-m)^2}{(n-b-m)^2}\cdot\frac{\sigma^2}{n-m} + \frac{2b^2}{(n-b-m)^2}\cdot\frac{(n-m)\sigma^2}{b} \\
&= \frac{2(n-m)\sigma^2}{(n-b-m)^2} + \frac{2b(n-m)\sigma^2}{(n-b-m)^2} \\
&= \frac{2(n-m)(b+1)\sigma^2}{(n-b-m)^2}.
\end{aligned}$$

Putting all the above components together, one has the following:

$$\mathbb{E}\left\|\frac{1}{|\mathcal{G}'|}\sum_{\boldsymbol{g}\in\mathcal{G}'}\theta_{\boldsymbol{g}}-\omega\right\|_2^2 \leq \frac{2(n-m)(b+1)\sigma^2}{(n-b-m)^2}.$$

The proof is complete.

B. Dataset Description

Detailed descriptions of the datasets used to evaluate our attack and defense method are provided below. **Texas100 [1]:** This dataset comprises hospital discharge records, containing inpatient data from various medical facilities, as published by the Texas Department of State Health Services. It includes 67,330 records with 6,170 binary features representing the 100 most frequently performed medical procedures. The records are organized into 100 distinct categories, each representing a unique patient type.

CIFAR-10 [28]: This dataset is a well-established benchmark for real-world object recognition, comprising 60,000 color images distributed evenly across 10 classes. It includes 50,000 images for training and 10,000 for testing, with a balanced number of images in each class.

STL10 [12]: Like CIFAR-10, this dataset is designed for image recognition and includes 10 classes, with 5,000 labeled images for training and 8,000 images for testing.

FER2013 [21]: This dataset consists of 35,886 grayscale images depicting facial expressions, divided into 28,708 training images, 3,589 PublicTest images, and 3,589 PrivateTest images. The images represent seven expression categories: anger, disgust, fear, happiness, sadness, surprise, and neutral.

C. Attack Description

Passive Membership Inference Attack [42]: Once the global model is downloaded from the server, the attacker determines an input sample to be a member if the model predicts it correctly; otherwise, it is classified as a non-member.

Gradient Ascent (GA) [36]: The attack uses gradient ascent on target samples to heighten the prediction gap between members and non-members. Upon receiving the global model parameters, it conducts inference in the manner of a passive membership inference attack.

AGREvader [57]: Rather than only altering the attack samples, AGREvader blends the attack gradients with normal gradients to ensure that the resulting combined gradients remain close to benign gradients in Euclidean norm, preventing noticeable deviation.

Adaptive attack: We examine a strong adversarial setting where the attacker is fully aware of the server's use of ATM. In this scenario, an adaptive attack is devised by carefully constructing gradients that inherently evade ATM filtering. The pseudocode for this attack is presented in Algorithm 3.

D. Comparison Defenses

We evaluate the performance of our attack and defense using the following mechanisms:

Differential Privacy [13]: The server adds Gaussian noise to all received gradients before performing the aggregation operation.

Top-k [2]: This approach selects the top k gradient dimensions with the highest absolute values for updates in the aggregation process, setting all other dimensions to zero.

FedAvg [34]: This trivial aggregation rule takes a simple average of the client updates.

Median [55]: This method computes the element-wise median of the gradients in the set $\mathcal{G}$, where $\mathcal{G}$ denotes all clients' uploaded gradients.

Trimmed-mean [55]: Once the server receives the set of all selected update gradients $\mathcal{G}$, for each dimension, it removes the largest b and smallest b elements before calculating their average.

Multi-Krum [7]: Upon receiving each model update, the server begins by identifying the $n - f - 1$ updates that are closest in terms of Euclidean distance, where f is the number of malicious clients. It then computes a cumulative score by aggregating these nearby updates. The update with the lowest calculated score is subsequently added to a candidate set. This selection and scoring process is repeated iteratively until a total of k updates have been selected. Once the candidate set is complete, the server updates the global model by aggregating all the chosen candidate updates. This method ensures that only the most consistent and reliable updates contribute to the global model, enhancing the robustness and accuracy of the federated learning system.

Fang [14]: The Fang defense method utilizes two techniques: Error Rate Rejection (ERR) and Loss Function Rejection (LFR), to filter out gradients from potentially malicious participants. By removing gradients that most negatively affect the error rate and loss, respectively, these methods strengthen the model's robustness. This selective exclusion helps ensure that only gradients that contribute positively to the model's performance are retained, improving its overall resilience against adversarial influences.

DeepSight [38]: The mechanism begins by calculating division differences and normalized update energies, then clusters the update gradients based on these metrics and cosine similarity. The cluster labels are refined through a voting scheme. Afterward, ℓ_2-norm clipping is applied to each benign gradient, and the clipped gradients are aggregated to update the global model. This process ensures that only reliable gradients contribute to the model update, enhancing the system's robustness.

E. Parameters Setting

In the default FL training scenario, there are 10 clients in total, consisting of both benign and malicious clients, with

Algorithm 2 ATM algorithm.

Input: Gradients from n clients: $\mathcal{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_n\}$, trim parameter b.
Output: Aggregated gradient $\bar{\boldsymbol{g}}$.

Initialize an $n \times n$ zero matrix $\boldsymbol{A}$ to record the angles between gradients.
for each gradient pair (i, j) where $1 \leq i < j \leq n$ **do**
    $\theta_{i,j} = arccos(\frac{\boldsymbol{g}_i \cdot \boldsymbol{g}_j}{\|\boldsymbol{g}_i\| \|\boldsymbol{g}_j\|})$
    $\boldsymbol{A}[i, j] \leftarrow \theta_{i,j}$
end for
for each row in $\boldsymbol{A}$ **do**
    $\bar{\theta}_i = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{A}[i, j]$
end for
Discard the $2b$ gradients with the largest absolute values in $\bar{\theta}$; denote the remaining set as $\hat{\mathcal{G}}$.
Calculate $\bar{g}$ by averaging the selected gradients as $\bar{\boldsymbol{g}} = \frac{1}{|\hat{\mathcal{G}}|} \sum_{\boldsymbol{g} \in \hat{\mathcal{G}}} \boldsymbol{g}$
Send aggregated gradient $\bar{\boldsymbol{g}}$ to clients.

10% of the clients being malicious and conducting the full-knowledge attack. In the partial-knowledge attack scenario, there are 50 clients in total, with 10 clients being malicious. During each training round, we assume that 80% of the clients participate in the training process. The attacker possesses 300 attack samples ($|D_{\text{attack}}|$) and 300 masking samples ($|D_{\text{mask}}|$). By default, we set $\gamma = 0.1$ when constructing $\hat{D}_{\text{mask}}$. To account for the worst-case scenario, we assume the attacker begins launching their attack in the first training round. For model training, we utilized the ResNet-20 [28] architecture on the CIFAR-10, STL10, and FER2013 datasets, while employing fully connected models for the Texas100 dataset. For all datasets, we set the training duration to 800 epochs, with a batch size of 64 and a learning rate of 0.01. To optimize the model, we used the Adam optimizer [25], which dynamically adjusts the learning rate, momentum, and other training parameters throughout the training process.

Algorithm 3 Adaptive attack against ATM.

Input: Total number of clients n, a set of benign gradients $\mathcal{G}_{\mathcal{B}}$, attack gradient $\boldsymbol{g}_{\text{attack}}$, trim parameter b.
Output: Adjusted malicious gradient $\boldsymbol{g}_{\text{adaptive}}$.

repeat
    Initialize an $n \times n$ zero matrix $\boldsymbol{A}$ to record the angles between gradients.
    for gradient pair (i, j) where $1 \leq i < j \leq n$ **do**
        $\theta_{i,j} = arccos(\frac{\boldsymbol{g}_i \cdot \boldsymbol{g}_j}{\|\boldsymbol{g}_i\| \|\boldsymbol{g}_j\|})$
        $\boldsymbol{A}[i, j] \leftarrow \theta_{i,j}$, $\boldsymbol{A}[j, i] \leftarrow \theta_{i,j}$
    end for
    for each row in $\boldsymbol{A}$ **do**
        $\bar{\theta}_i = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{A}[i, j]$
    end for
    Let θ_{attack} represent the final value of $\bar{\theta}$, corresponding to the average angular deviation between the attack gradient and the benign gradients.
    Arrange $\bar{\theta}$ in ascending order and define the trimming threshold θ_τ as the angle ranked $2b$-th from the largest in the sorted list.
    if $\theta_{\text{attack}} < \theta_\tau$ **then**
        break ▷ Attack is considered successful
    else
        Select the benign gradient $\boldsymbol{g}_k$ with the greatest angular deviation from $\boldsymbol{g}_{\text{attack}}$.
        Update $\boldsymbol{g}_{\text{attack}} \leftarrow \frac{1}{2}(\boldsymbol{g}_{\text{attack}} + \boldsymbol{g}_k)$
    end if
until convergence
return $\boldsymbol{g}_{\text{adaptive}} \leftarrow \boldsymbol{g}_{\text{attack}}$

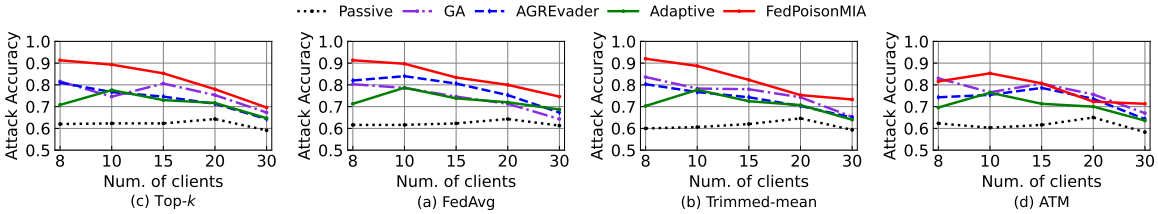

Figure 4. Impact of total number of clients.

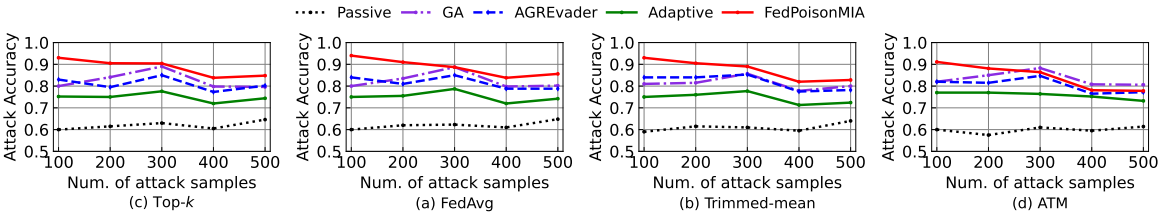

Figure 5. Impact of number of target samples.

Table 6. Attack accuracy with $C = 1.0$ in synchronous setting, where C represents the proportion of clients selected in each round.

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-Mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.643	0.623	0.646	0.620	0.650	0.623	0.560	0.550	0.633	0.606	0.616	0.596
	GA	0.823	0.790	0.813	0.783	0.816	0.780	0.736	0.720	0.813	0.756	0.800	0.596
	AGREvader	0.743	0.756	0.741	0.777	0.830	0.850	0.803	0.813	0.757	0.771	0.751	0.764
	FedPoisonMIA	0.894	0.914	0.893	0.906	0.893	0.900	0.923	0.950	0.887	0.890	0.834	0.883
CIFAR-10	Passive	0.607	0.603	0.600	0.590	0.580	0.600	0.567	0.570	0.577	0.563	0.583	0.583
	GA	0.737	0.603	0.733	0.610	0.723	0.630	0.693	0.590	0.707	0.597	0.710	0.603
	AGREvader	0.654	0.684	0.641	0.681	0.710	0.653	0.713	0.650	0.730	0.647	0.653	0.627
	FedPoisonMIA	0.907	0.783	0.853	0.777	0.847	0.780	0.783	0.700	0.910	0.790	0.783	0.698
STL10	Passive	0.600	0.570	0.557	0.563	0.610	0.580	0.547	0.543	0.570	0.560	0.587	0.587
	GA	0.800	0.697	0.803	0.730	0.800	0.720	0.737	0.657	0.757	0.683	0.786	0.757
	AGREvader	0.751	0.743	0.714	0.750	0.760	0.713	0.827	0.720	0.817	0.727	0.763	0.703
	FedPoisonMIA	0.877	0.803	0.863	0.797	0.833	0.807	0.827	0.763	0.827	0.773	0.823	0.730
FER2013	Passive	0.606	0.573	0.590	0.580	0.613	0.603	0.550	0.513	0.580	0.553	0.600	0.556
	GA	0.810	0.720	0.760	0.713	0.833	0.710	0.670	0.663	0.823	0.776	0.813	0.750
	AGREvader	0.870	0.793	0.790	0.790	0.860	0.820	0.743	0.746	0.810	0.760	0.728	0.734
	FedPoisonMIA	0.936	0.913	0.913	0.880	0.926	0.920	0.813	0.776	0.936	0.886	0.834	0.816

Table 7. Attack accuracy results in asynchronous setting.

(a) $C = 0.8$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.653	0.636	0.660	0.640	0.663	0.633	0.610	0.613	0.650	0.620	0.646	0.640
	GA	0.790	0.750	0.790	0.740	0.790	0.763	0.733	0.750	0.786	0.743	0.630	0.640
	AGREvader	0.743	0.726	0.736	0.753	0.803	0.820	0.748	0.760	0.746	0.740	0.726	0.740
	FedPoisonMIA	0.843	0.867	0.871	0.881	0.863	0.863	0.804	0.853	0.806	0.843	0.736	0.750
CIFAR-10	Passive	0.597	0.580	0.590	0.580	0.633	0.620	0.607	0.597	0.613	0.607	0.633	0.593
	GA	0.630	0.563	0.603	0.577	0.653	0.580	0.630	0.557	0.633	0.580	0.583	0.577
	AGREvader	0.603	0.567	0.627	0.560	0.660	0.673	0.638	0.617	0.645	0.661	0.633	0.597
	FedPoisonMIA	0.873	0.687	0.867	0.717	0.880	0.680	0.677	0.647	0.787	0.720	0.630	0.587
STL10	Passive	0.543	0.546	0.550	0.543	0.620	0.613	0.617	0.573	0.620	0.610	0.610	0.563
	GA	0.633	0.580	0.703	0.637	0.613	0.593	0.580	0.547	0.653	0.613	0.587	0.563
	AGREvader	0.537	0.520	0.543	0.557	0.667	0.673	0.686	0.603	0.656	0.720	0.557	0.560
	FedPoisonMIA	0.880	0.830	0.897	0.890	0.873	0.860	0.737	0.610	0.713	0.830	0.613	0.563
FER2013	Passive	0.626	0.616	0.610	0.623	0.650	0.610	0.590	0.550	0.606	0.596	0.581	0.570
	GA	0.710	0.606	0.756	0.706	0.716	0.660	0.673	0.573	0.640	0.643	0.736	0.687
	AGREvader	0.690	0.643	0.713	0.690	0.860	0.833	0.736	0.730	0.747	0.724	0.618	0.590
	FedPoisonMIA	0.900	0.870	0.926	0.870	0.920	0.903	0.766	0.763	0.943	0.910	0.726	0.680

(b) $C = 1.0$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.670	0.653	0.673	0.656	0.670	0.650	0.593	0.600	0.650	0.633	0.600	0.660
	GA	0.806	0.786	0.806	0.783	0.803	0.776	0.746	0.683	0.773	0.770	0.613	0.683
	AGREvader	0.750	0.736	0.743	0.740	0.794	0.850	0.757	0.760	0.750	0.751	0.739	0.746
	FedPoisonMIA	0.877	0.900	0.881	0.897	0.887	0.904	0.903	0.930	0.817	0.877	0.733	0.743
CIFAR-10	Passive	0.620	0.577	0.593	0.570	0.613	0.593	0.587	0.587	0.597	0.587	0.620	0.587
	GA	0.637	0.597	0.653	0.590	0.657	0.570	0.660	0.590	0.670	0.587	0.587	0.550
	AGREvader	0.607	0.573	0.623	0.563	0.650	0.667	0.657	0.660	0.651	0.667	0.643	0.586
	FedPoisonMIA	0.887	0.720	0.920	0.723	0.880	0.683	0.723	0.647	0.840	0.747	0.646	0.583
STL10	Passive	0.543	0.557	0.560	0.553	0.610	0.613	0.593	0.560	0.613	0.600	0.610	0.573
	GA	0.627	0.610	0.680	0.653	0.653	0.620	0.613	0.590	0.643	0.610	0.563	0.553
	AGREvader	0.560	0.550	0.553	0.543	0.663	0.680	0.663	0.687	0.663	0.683	0.597	0.550
	FedPoisonMIA	0.903	0.853	0.883	0.853	0.897	0.830	0.763	0.690	0.690	0.820	0.620	0.560
FER2013	Passive	0.626	0.603	0.623	0.593	0.620	0.613	0.573	0.556	0.606	0.586	0.570	0.566
	GA	0.750	0.640	0.726	0.690	0.730	0.700	0.653	0.603	0.656	0.656	0.683	0.660
	AGREvader	0.716	0.653	0.716	0.683	0.823	0.843	0.721	0.726	0.714	0.730	0.724	0.633
	FedPoisonMIA	0.890	0.866	0.916	0.870	0.887	0.923	0.856	0.740	0.920	0.940	0.741	0.696

Table 8. Attack precision results in synchronous setting.

(a) $C = 0.8$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.593	0.579	0.590	0.577	0.578	0.565	0.553	0.542	0.578	0.565	0.559	0.568
	GA	0.750	0.711	0.748	0.679	0.750	0.736	0.767	0.755	0.741	0.725	0.680	0.707
	AGREvader	0.685	0.684	0.685	0.683	0.719	0.799	0.677	0.668	0.674	0.683	0.659	0.690
	FedPoisonMIA	0.821	0.825	0.821	0.844	0.830	0.830	0.853	0.891	0.807	0.816	0.719	0.774
CIFAR-10	Passive	0.548	0.547	0.539	0.544	0.545	0.539	0.583	0.598	0.564	0.562	0.579	0.559
	GA	0.633	0.609	0.649	0.585	0.659	0.583	0.660	0.590	0.659	0.624	0.590	0.585
	AGREvader	0.590	0.608	0.584	0.605	0.629	0.578	0.665	0.567	0.642	0.603	0.575	0.596
	FedPoisonMIA	0.913	0.722	0.927	0.790	0.919	0.779	0.704	0.612	0.783	0.837	0.645	0.589
STL10	Passive	0.563	0.564	0.565	0.555	0.570	0.556	0.543	0.541	0.552	0.543	0.577	0.553
	GA	0.770	0.732	0.768	0.743	0.767	0.748	0.794	0.727	0.781	0.742	0.653	0.633
	AGREvader	0.651	0.649	0.619	0.674	0.787	0.704	0.724	0.696	0.789	0.780	0.618	0.628
	FedPoisonMIA	0.914	0.792	0.931	0.840	0.890	0.795	0.801	0.746	0.769	0.847	0.753	0.692
FER2013	Passive	0.589	0.568	0.581	0.579	0.566	0.541	0.555	0.521	0.566	0.541	0.580	0.631
	GA	0.751	0.727	0.724	0.728	0.746	0.713	0.738	0.740	0.764	0.743	0.655	0.665
	AGREvader	0.845	0.765	0.701	0.788	0.877	0.868	0.753	0.667	0.868	0.850	0.681	0.670
	FedPoisonMIA	0.918	0.887	0.917	0.874	0.922	0.923	0.806	0.790	0.928	0.925	0.707	0.728

(b) $C = 1.0$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.587	0.574	0.589	0.578	0.583	0.566	0.535	0.533	0.583	0.566	0.570	0.559
	GA	0.754	0.743	0.745	0.732	0.749	0.728	0.755	0.775	0.749	0.733	0.709	0.719
	AGREvader	0.664	0.682	0.659	0.693	0.751	0.810	0.722	0.683	0.674	0.686	0.668	0.680
	FedPoisonMIA	0.825	0.844	0.825	0.853	0.825	0.834	0.868	0.919	0.816	0.821	0.751	0.812
CIFAR-10	Passive	0.558	0.538	0.560	0.568	0.547	0.600	0.573	0.590	0.543	0.576	0.547	0.549
	GA	0.662	0.585	0.662	0.586	0.654	0.595	0.683	0.597	0.646	0.574	0.638	0.589
	AGREvader	0.592	0.614	0.583	0.611	0.638	0.610	0.698	0.669	0.656	0.609	0.591	0.591
	FedPoisonMIA	0.918	0.732	0.831	0.727	0.851	0.747	0.801	0.703	0.856	0.774	0.846	0.624
STL10	Passive	0.563	0.566	0.559	0.557	0.567	0.547	0.533	0.525	0.543	0.534	0.557	0.552
	GA	0.771	0.712	0.751	0.756	0.771	0.770	0.775	0.790	0.765	0.743	0.715	0.758
	AGREvader	0.668	0.664	0.637	0.668	0.717	0.691	0.799	0.770	0.781	0.705	0.685	0.667
	FedPoisonMIA	0.918	0.749	0.847	0.795	0.809	0.819	0.761	0.762	0.758	0.744	0.757	0.682
FER2013	Passive	0.564	0.543	0.552	0.548	0.548	0.533	0.535	0.508	0.548	0.533	0.556	0.536
	GA	0.757	0.712	0.722	0.703	0.794	0.717	0.758	0.788	0.790	0.786	0.735	0.717
	AGREvader	0.873	0.828	0.740	0.800	0.923	0.839	0.662	0.667	0.904	0.791	0.648	0.654
	FedPoisonMIA	0.917	0.919	0.887	0.880	0.900	0.938	0.952	0.855	0.934	0.926	0.751	0.891

Table 9. Attack precision results in asynchronous setting.

(a) $C = 0.8$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.596	0.595	0.600	0.594	0.590	0.579	0.583	0.615	0.590	0.579	0.608	0.648
	GA	0.723	0.702	0.719	0.688	0.719	0.691	0.711	0.727	0.724	0.709	0.627	0.618
	AGREvader	0.680	0.660	0.664	0.684	0.738	0.740	0.653	0.673	0.675	0.661	0.675	0.671
	FedPoisonMIA	0.764	0.791	0.795	0.807	0.786	0.791	0.719	0.774	0.723	0.763	0.656	0.673
CIFAR-10	Passive	0.566	0.561	0.557	0.603	0.583	0.571	0.565	0.590	0.566	0.567	0.578	0.608
	GA	0.584	0.574	0.568	0.591	0.602	0.560	0.601	0.584	0.588	0.562	0.574	0.577
	AGREvader	0.579	0.582	0.577	0.541	0.596	0.613	0.581	0.575	0.585	0.597	0.579	0.587
	FedPoisonMIA	0.818	0.624	0.820	0.684	0.817	0.638	0.611	0.616	0.705	0.699	0.577	0.573
STL10	Passive	0.533	0.546	0.535	0.550	0.571	0.574	0.574	0.560	0.583	0.574	0.585	0.627
	GA	0.585	0.582	0.652	0.610	0.582	0.577	0.570	0.554	0.609	0.585	0.641	0.661
	AGREvader	0.538	0.524	0.533	0.583	0.612	0.616	0.626	0.591	0.600	0.642	0.559	0.632
	FedPoisonMIA	0.828	0.766	0.848	0.842	0.818	0.814	0.662	0.586	0.637	0.749	0.573	0.564
FER2013	Passive	0.581	0.576	0.569	0.611	0.570	0.561	0.573	0.540	0.570	0.561	0.563	0.577
	GA	0.663	0.607	0.708	0.605	0.672	0.621	0.638	0.571	0.706	0.611	0.668	0.658
	AGREvader	0.634	0.582	0.654	0.600	0.800	0.794	0.668	0.665	0.665	0.645	0.613	0.597
	FedPoisonMIA	0.857	0.807	0.890	0.803	0.889	0.854	0.698	0.663	0.935	0.897	0.650	0.627

(b) $C = 1.0$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.606	0.609	0.609	0.611	0.593	0.590	0.592	0.570	0.593	0.590	0.585	0.617
	GA	0.730	0.729	0.730	0.730	0.726	0.722	0.731	0.695	0.703	0.734	0.580	0.628
	AGREvader	0.676	0.670	0.670	0.673	0.709	0.772	0.674	0.679	0.671	0.668	0.651	0.674
	FedPoisonMIA	0.803	0.834	0.807	0.830	0.816	0.803	0.831	0.878	0.733	0.803	0.675	0.665
CIFAR-10	Passive	0.583	0.564	0.556	0.571	0.569	0.562	0.583	0.603	0.563	0.551	0.569	0.584
	GA	0.589	0.565	0.599	0.562	0.604	0.547	0.624	0.569	0.615	0.559	0.571	0.566
	AGREvader	0.565	0.553	0.575	0.543	0.590	0.606	0.594	0.603	0.590	0.604	0.599	0.557
	FedPoisonMIA	0.858	0.650	0.884	0.656	0.875	0.624	0.660	0.607	0.760	0.728	0.590	0.619
STL10	Passive	0.536	0.559	0.541	0.545	0.569	0.571	0.558	0.549	0.572	0.564	0.571	0.608
	GA	0.591	0.582	0.627	0.616	0.612	0.603	0.588	0.583	0.614	0.587	0.600	0.593
	AGREvader	0.574	0.551	0.541	0.547	0.599	0.611	0.598	0.623	0.598	0.613	0.571	0.600
	FedPoisonMIA	0.850	0.791	0.840	0.827	0.848	0.762	0.680	0.636	0.653	0.740	0.576	0.587
FER2013	Passive	0.590	0.577	0.579	0.559	0.573	0.552	0.552	0.536	0.573	0.552	0.598	0.585
	GA	0.705	0.647	0.687	0.676	0.686	0.662	0.682	0.645	0.683	0.654	0.579	0.628
	AGREvader	0.631	0.595	0.641	0.611	0.741	0.805	0.643	0.653	0.637	0.663	0.645	0.601
	FedPoisonMIA	0.834	0.812	0.879	0.828	0.887	0.822	0.809	0.712	0.923	0.909	0.659	0.628

Table 10. Attack recall results in synchronous setting.

(a) $C = 0.8$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.960	0.927	0.960	0.920	0.960	0.933	0.867	0.853	0.960	0.933	0.940	0.860
	GA	0.980	0.953	0.987	0.947	0.980	0.893	0.920	0.800	0.953	0.913	0.893	0.933
	AGREvader	0.987	0.953	0.987	1.000	1.000	1.000	1.000	0.993	0.993	1.000	1.000	1.000
	FedPoisonMIA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.980	1.000	1.000	1.000	1.000
CIFAR-10	Passive	0.993	0.927	0.960	0.873	0.960	0.960	0.420	0.447	0.527	0.540	1.000	0.853
	GA	0.933	0.780	0.927	0.780	0.940	0.773	0.907	0.633	0.940	0.787	1.000	0.847
	AGREvader	1.000	0.993	0.993	0.980	0.973	0.840	0.820	0.767	0.920	0.800	0.973	0.893
	FedPoisonMIA	0.913	0.900	0.927	0.727	0.907	0.800	0.873	0.873	0.987	0.753	0.947	0.993
STL10	Passive	0.927	0.913	0.953	0.840	0.947	0.927	0.927	0.613	0.913	0.920	0.853	0.560
	GA	0.913	0.673	0.947	0.713	0.880	0.673	0.773	0.480	0.833	0.593	0.980	0.840
	AGREvader	0.993	0.987	1.000	0.980	0.813	0.807	0.893	0.747	0.847	0.687	0.980	0.833
	FedPoisonMIA	0.927	0.913	0.900	0.807	0.913	0.853	0.967	0.860	0.953	0.847	0.973	0.987
FER2013	Passive	0.927	0.807	0.933	0.933	0.887	0.880	0.873	0.893	0.887	0.880	0.920	0.900
	GA	0.907	0.960	0.893	0.947	0.980	0.893	0.733	0.607	0.973	0.887	0.987	0.860
	AGREvader	0.833	0.760	0.920	0.767	0.953	0.833	0.813	0.987	0.833	0.793	0.970	0.960
	FedPoisonMIA	0.967	0.940	0.960	0.927	0.947	0.953	0.833	0.727	0.947	0.900	0.980	0.980

(b) $C = 1.0$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.967	0.953	0.967	0.893	0.933	0.913	0.913	0.813	0.933	0.913	0.947	0.913
	GA	0.960	0.887	0.953	0.893	0.953	0.893	0.700	0.620	0.953	0.807	0.927	0.763
	AGREvader	0.987	0.960	1.000	1.000	1.000	0.993	0.987	0.833	1.000	1.000	1.000	1.000
	FedPoisonMIA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.987	1.000	1.000	1.000	1.000
CIFAR-10	Passive	0.960	0.987	0.527	0.527	0.927	0.600	0.527	0.460	0.960	0.480	0.973	0.940
	GA	0.967	0.713	0.953	0.747	0.947	0.813	0.720	0.553	0.913	0.747	0.973	0.687
	AGREvader	1.000	1.000	1.000	1.000	0.973	0.853	0.753	0.593	0.967	0.820	0.993	0.820
	FedPoisonMIA	0.893	0.893	0.887	0.887	0.840	0.847	0.753	0.693	0.987	0.820	0.693	1.000
STL10	Passive	0.953	0.887	0.940	0.880	0.933	0.933	0.753	0.900	0.887	0.933	0.840	0.913
	GA	0.853	0.660	0.907	0.680	0.853	0.627	0.667	0.427	0.740	0.560	0.953	0.753
	AGREvader	1.000	0.987	1.000	0.993	0.860	0.773	0.873	0.627	0.880	0.780	0.973	0.813
	FedPoisonMIA	0.827	0.913	0.887	0.800	0.873	0.787	0.953	0.767	0.960	0.833	0.953	0.860
FER2013	Passive	0.940	0.920	0.953	0.920	0.907	0.867	0.753	0.887	0.907	0.867	0.987	0.833
	GA	0.913	0.740	0.847	0.740	0.900	0.693	0.500	0.447	0.880	0.760	0.980	0.827
	AGREvader	0.867	0.740	0.893	0.773	0.800	0.693	0.993	0.987	0.693	0.707	1.000	1.000
	FedPoisonMIA	0.960	0.907	0.947	0.880	0.960	0.900	0.660	0.667	0.940	0.840	1.000	0.927

Table 11. Attack recall results in asynchronous setting.

(a) $C = 0.8$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.953	0.853	0.960	0.887	0.980	0.880	0.773	0.607	0.980	0.880	0.827	0.613
	GA	0.940	0.893	0.953	0.880	0.953	0.880	0.920	0.800	0.927	0.827	0.640	0.733
	AGREvader	0.920	0.933	0.960	0.940	0.940	0.967	0.921	0.987	0.873	0.987	0.873	0.940
	FedPoisonMIA	0.993	1.000	1.000	1.000	1.000	0.987	1.000	1.000	0.993	1.000	0.993	0.973
CIFAR-10	Passive	0.827	0.740	0.880	0.467	0.940	0.960	0.933	0.633	0.973	0.900	0.987	0.527
	GA	0.900	0.467	0.867	0.500	0.907	0.747	0.773	0.393	0.893	0.727	0.647	0.573
	AGREvader	0.760	0.473	0.953	0.793	0.993	0.940	1.000	0.900	1.000	1.000	0.973	0.653
	FedPoisonMIA	0.960	0.940	0.940	0.807	0.980	0.833	0.973	0.780	0.987	0.773	0.980	0.680
STL10	Passive	0.693	0.553	0.767	0.480	0.960	0.880	0.907	0.680	0.840	0.853	0.760	0.313
	GA	0.920	0.567	0.873	0.760	0.807	0.700	0.653	0.480	0.860	0.780	0.393	0.260
	AGREvader	0.520	0.440	0.700	0.400	0.913	0.920	0.927	0.673	0.940	0.993	0.533	0.287
	FedPoisonMIA	0.960	0.960	0.967	0.960	0.960	0.933	0.967	0.747	0.993	0.993	0.887	0.560
FER2013	Passive	0.913	0.880	0.907	0.680	0.867	0.887	0.707	0.673	0.867	0.887	0.857	0.837
	GA	0.853	0.607	0.873	0.847	0.847	0.720	0.800	0.587	0.833	0.787	0.940	0.720
	AGREvader	0.900	0.760	0.907	0.700	0.960	0.900	0.980	0.993	1.000	1.000	0.940	0.933
	FedPoisonMIA	0.960	0.973	0.973	0.980	0.960	0.973	0.940	0.747	0.953	0.927	0.980	0.887

(b) $C = 1.0$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.973	0.860	0.967	0.860	0.953	0.873	0.600	0.813	0.953	0.873	0.687	0.880
	GA	0.973	0.913	0.973	0.900	0.973	0.900	0.780	0.653	0.947	0.847	0.820	0.900
	AGREvader	0.960	0.933	0.960	0.933	1.000	0.993	1.000	0.987	0.980	1.000	0.957	0.953
	FedPoisonMIA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.900	0.980
CIFAR-10	Passive	0.840	0.673	0.933	0.560	0.940	0.847	0.607	0.507	0.860	0.940	0.993	0.600
	GA	0.907	0.840	0.927	0.820	0.907	0.813	0.807	0.747	0.907	0.820	0.700	0.427
	AGREvader	0.927	0.767	0.947	0.800	0.987	0.953	0.993	0.940	1.000	0.967	0.867	0.840
	FedPoisonMIA	0.927	0.953	0.967	0.940	0.887	0.920	0.920	0.833	0.993	0.787	0.960	0.433
STL10	Passive	0.640	0.533	0.793	0.647	0.913	0.913	0.900	0.667	0.900	0.880	0.880	0.413
	GA	0.820	0.780	0.887	0.813	0.840	0.700	0.760	0.633	0.773	0.740	0.380	0.340
	AGREvader	0.467	0.540	0.707	0.507	0.987	0.993	0.993	0.947	0.993	0.993	0.780	0.300
	FedPoisonMIA	0.980	0.960	0.947	0.893	0.967	0.960	0.993	0.887	0.993	0.987	0.907	0.407
FER2013	Passive	0.833	0.773	0.900	0.880	0.833	0.927	0.773	0.840	0.833	0.927	0.803	0.781
	GA	0.860	0.660	0.833	0.653	0.847	0.653	0.573	0.460	0.747	0.667	0.660	0.787
	AGREvader	0.900	0.753	0.987	0.773	0.993	0.907	1.000	0.967	1.000	0.933	1.000	0.793
	FedPoisonMIA	0.973	0.953	0.967	0.933	0.993	0.987	0.933	0.807	0.960	0.933	1.000	0.967

Table 12. Test accuracy results in synchronous setting.

(a) $C = 0.8$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.577	0.570	0.577	0.569	0.579	0.571	0.583	0.570	0.579	0.571	0.583	0.566
	GA	0.572	0.565	0.573	0.573	0.573	0.566	0.572	0.563	0.571	0.566	0.578	0.567
	AGREvader	0.576	0.567	0.575	0.574	0.572	0.568	0.574	0.558	0.578	0.570	0.580	0.568
	FedPoisonMIA	0.577	0.565	0.573	0.562	0.575	0.563	0.566	0.544	0.571	0.555	0.576	0.569
CIFAR-10	Passive	0.790	0.741	0.766	0.741	0.765	0.741	0.727	0.677	0.768	0.740	0.785	0.754
	GA	0.763	0.730	0.765	0.728	0.765	0.728	0.738	0.716	0.760	0.729	0.766	0.719
	AGREvader	0.760	0.742	0.763	0.742	0.768	0.725	0.741	0.705	0.755	0.731	0.773	0.729
	FedPoisonMIA	0.757	0.725	0.750	0.723	0.750	0.730	0.722	0.677	0.755	0.706	0.748	0.775
STL10	Passive	0.593	0.541	0.548	0.513	0.545	0.508	0.560	0.529	0.563	0.502	0.565	0.532
	GA	0.531	0.472	0.536	0.473	0.536	0.476	0.533	0.495	0.528	0.475	0.591	0.545
	AGREvader	0.578	0.524	0.551	0.497	0.540	0.489	0.549	0.499	0.539	0.498	0.581	0.535
	FedPoisonMIA	0.554	0.468	0.503	0.471	0.516	0.486	0.565	0.475	0.548	0.467	0.581	0.594
FER2013	Passive	0.586	0.555	0.572	0.549	0.563	0.550	0.560	0.548	0.563	0.550	0.582	0.547
	GA	0.540	0.524	0.552	0.533	0.536	0.519	0.513	0.509	0.535	0.517	0.565	0.527
	AGREvader	0.531	0.510	0.556	0.519	0.542	0.525	0.513	0.529	0.545	0.519	0.561	0.552
	FedPoisonMIA	0.546	0.525	0.543	0.548	0.546	0.528	0.489	0.547	0.530	0.514	0.570	0.539

(b) $C = 1.0$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.575	0.571	0.575	0.569	0.577	0.570	0.583	0.572	0.577	0.570	0.580	0.575
	GA	0.572	0.567	0.571	0.567	0.572	0.565	0.573	0.558	0.572	0.565	0.573	0.563
	AGREvader	0.574	0.569	0.578	0.568	0.570	0.569	0.571	0.568	0.577	0.567	0.579	0.569
	FedPoisonMIA	0.571	0.562	0.575	0.560	0.570	0.561	0.563	0.541	0.569	0.554	0.575	0.563
CIFAR-10	Passive	0.766	0.736	0.772	0.732	0.772	0.732	0.756	0.685	0.771	0.721	0.774	0.750
	GA	0.755	0.730	0.758	0.735	0.767	0.733	0.739	0.714	0.756	0.726	0.756	0.719
	AGREvader	0.760	0.737	0.761	0.735	0.761	0.721	0.736	0.706	0.759	0.731	0.776	0.731
	FedPoisonMIA	0.748	0.730	0.742	0.707	0.739	0.722	0.719	0.673	0.743	0.717	0.731	0.783
STL10	Passive	0.576	0.532	0.546	0.486	0.555	0.483	0.552	0.511	0.549	0.502	0.619	0.516
	GA	0.522	0.475	0.522	0.464	0.533	0.468	0.528	0.455	0.529	0.459	0.564	0.512
	AGREvader	0.579	0.509	0.547	0.470	0.561	0.463	0.542	0.489	0.533	0.473	0.564	0.519
	FedPoisonMIA	0.520	0.483	0.528	0.469	0.505	0.463	0.528	0.453	0.508	0.440	0.582	0.499
FER2013	Passive	0.568	0.557	0.567	0.552	0.566	0.561	0.556	0.546	0.566	0.561	0.583	0.556
	GA	0.537	0.525	0.551	0.528	0.536	0.530	0.507	0.504	0.513	0.519	0.544	0.519
	AGREvader	0.548	0.533	0.538	0.534	0.539	0.518	0.546	0.545	0.540	0.512	0.548	0.538
	FedPoisonMIA	0.540	0.515	0.539	0.545	0.533	0.519	0.515	0.507	0.532	0.519	0.548	0.537

Table 13. Test accuracy results in asynchronous setting.

(a) $C = 0.8$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.570	0.572	0.567	0.570	0.575	0.572	0.588	0.566	0.575	0.572	0.552	0.516
	GA	0.569	0.566	0.569	0.567	0.570	0.566	0.583	0.563	0.571	0.565	0.511	0.465
	AGREvader	0.565	0.575	0.561	0.571	0.584	0.578	0.581	0.571	0.579	0.587	0.579	0.476
	FedPoisonMIA	0.593	0.581	0.557	0.581	0.586	0.579	0.583	0.556	0.592	0.581	0.583	0.571
CIFAR-10	Passive	0.743	0.706	0.776	0.717	0.783	0.763	0.713	0.679	0.779	0.752	0.748	0.645
	GA	0.763	0.698	0.763	0.712	0.749	0.706	0.733	0.650	0.759	0.702	0.735	0.608
	AGREvader	0.723	0.707	0.764	0.710	0.784	0.764	0.790	0.773	0.793	0.779	0.796	0.763
	FedPoisonMIA	0.774	0.742	0.768	0.730	0.761	0.745	0.721	0.652	0.770	0.741	0.800	0.673
STL10	Passive	0.594	0.540	0.602	0.538	0.676	0.630	0.640	0.588	0.691	0.635	0.625	0.479
	GA	0.542	0.533	0.526	0.513	0.519	0.525	0.599	0.490	0.536	0.509	0.487	0.366
	AGREvader	0.576	0.527	0.601	0.491	0.679	0.623	0.558	0.511	0.670	0.640	0.637	0.453
	FedPoisonMIA	0.617	0.579	0.601	0.575	0.611	0.591	0.599	0.561	0.661	0.584	0.606	0.429
FER2013	Passive	0.550	0.553	0.552	0.553	0.578	0.561	0.552	0.556	0.578	0.561	0.529	0.528
	GA	0.539	0.532	0.543	0.526	0.553	0.544	0.535	0.531	0.549	0.532	0.508	0.578
	AGREvader	0.551	0.542	0.535	0.535	0.567	0.510	0.556	0.542	0.566	0.538	0.535	0.513
	FedPoisonMIA	0.559	0.531	0.555	0.542	0.557	0.545	0.563	0.542	0.534	0.528	0.558	0.538

(b) $C = 1.0$

Dataset	Attack	DP		Top-k		FedAvg		Median		Trimmed-mean		ATM	
		IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
Texas100	Passive	0.573	0.578	0.570	0.578	0.575	0.584	0.585	0.586	0.575	0.584	0.577	0.561
	GA	0.570	0.572	0.569	0.574	0.568	0.576	0.582	0.579	0.570	0.581	0.567	0.551
	AGREvader	0.559	0.568	0.558	0.570	0.573	0.576	0.592	0.575	0.599	0.588	0.591	0.581
	FedPoisonMIA	0.586	0.580	0.584	0.582	0.590	0.577	0.567	0.554	0.592	0.576	0.592	0.591
CIFAR-10	Passive	0.747	0.714	0.765	0.754	0.780	0.741	0.726	0.667	0.792	0.761	0.745	0.669
	GA	0.731	0.719	0.765	0.746	0.749	0.726	0.734	0.735	0.750	0.724	0.746	0.590
	AGREvader	0.723	0.704	0.759	0.719	0.788	0.711	0.780	0.760	0.780	0.756	0.721	0.674
	FedPoisonMIA	0.765	0.715	0.772	0.702	0.768	0.727	0.723	0.680	0.772	0.736	0.752	0.768
STL10	Passive	0.557	0.513	0.609	0.545	0.667	0.629	0.652	0.601	0.669	0.618	0.633	0.584
	GA	0.514	0.476	0.517	0.491	0.585	0.547	0.593	0.550	0.548	0.569	0.488	0.386
	AGREvader	0.569	0.544	0.598	0.538	0.681	0.633	0.670	0.620	0.664	0.631	0.553	0.542
	FedPoisonMIA	0.631	0.586	0.597	0.527	0.605	0.571	0.620	0.571	0.643	0.595	0.539	0.549
FER2013	Passive	0.557	0.554	0.571	0.537	0.568	0.550	0.562	0.538	0.568	0.550	0.532	0.526
	GA	0.550	0.529	0.536	0.536	0.543	0.538	0.542	0.517	0.547	0.536	0.549	0.519
	AGREvader	0.555	0.534	0.549	0.534	0.544	0.541	0.554	0.537	0.561	0.549	0.556	0.541
	FedPoisonMIA	0.538	0.518	0.527	0.536	0.552	0.539	0.501	0.537	0.546	0.536	0.552	0.517

Table 14. Attack running time (seconds).

Dataset	AGREvader	FedPoisonMIA
Texas100	1104.832	1669.040
CIFAR-10	3673.408	4404.120
STL10	3508.848	4316.448
FER2013	3173.304	3810.128

Table 15. Computational cost (in seconds) of different defenses.

Dataset	Median		Trimmed-mean		ATM	
	10 clients	50 clients	10 clients	50 clients	10 clients	50 clients
Texas100	5.262	18.344	2.942	20.944	0.975	4.688
CIFAR-10	0.320	0.618	0.250	1.729	0.295	0.371
STL10	0.312	0.700	0.250	1.658	0.314	0.405
FER2013	0.311	0.759	0.242	1.646	0.282	0.372