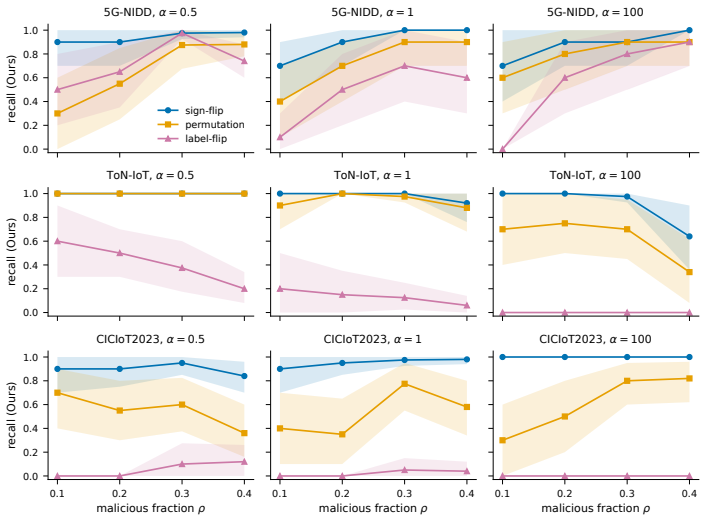

Fig. S1 Recall of the audit (Ours) on the confirmation split versus the malicious fraction ρ, with 95% bootstrap bands over the 10 seeds.

Online Resource 1. Supplementary tables and figures for "Calibrated, Feature-Attributed Accusation of Influential Response-Curve Edits in Label-Skewed Federated Intrusion Detection with Glass-Box Additive Models"

Abstract

This resource holds the secondary analyses that the article cites. Section numbers of the article are written as "article Section n". Every value comes from the reproducibility package of the article.

Table S1 Ablation of the null, with the cosine score: any-false-flag rate in clean federations (10 seeds per value) at $\alpha = 0.5$ / 1 / 100 on the exploration split, and sign-flip recall at $\rho = 0.2$. For MQTTset, the pooled any-false-flag rate and the range of recall over α.

	Any-false-flag		SF recall, $\rho = 0.2$		MQTTset
Null	5G-NIDD	ToN-IoT	5G-NIDD	ToN-IoT	any-ff / SF recall
Two-pass bootstrap	0.2 / 0.8 / 0.2	0.1 / 0.1 / 0.1	0.90 / 1.00 / 1.00	0.90 / 1.00 / 1.00	0.37 / 0.40–1.00
Leave-one-out t	0.1 / 0.8 / 0.1	0.1 / 0.1 / 0.1	0.80 / 1.00 / 1.00	0.90 / 1.00 / 1.00	0.20 / 0.30–0.90
Median and interquartile-range null	0.2 / 0.5 / 0.4	0.2 / 0.1 / 0.2	0.70 / 0.80 / 1.00	0.90 / 1.00 / 1.00	0.20 / 0.20–0.70
Simulation + core shift, $M = 32$	0.1 / 0.1 / 0.0	0.0 / 0.0 / 0.5	0.80 / 1.00 / 1.00	1.00 / 1.00 / 1.00	0.13 / 0.35–1.00
Simulation, empirical tail, $M = 48$	0 / 0 / 0	0 / 0 / 0	0.00 / 0.00 / 0.00	0.10 / 0.00 / 0.00	0.00 / 0.00–0.10
Simulation, log-t tail, $M = 48$ (Ours)	0 / 0 / 0	0 / 0 / 0	0.75 / 0.90 / 1.00	1.00 / 1.00 / 1.00	0.00 / 0.00–0.25

Table S2 Ablation of the other components (design set and exploration split, $\alpha = 0.5$). Rankings of the statistics are in Table S3. The interaction model is a generalized additive model with pairwise interactions fitted as a federated EBM.

Component	Setting	Effect
Statistic	sum-L_2, per-feature max-z, with the bootstrap null	sum-L_2 almost never accuses (sign-flip recall 0–0.2); per-feature has clean FPR 0.13–0.25
Grid size	$G \in \{10, 20, 40, 80, 160\}$	sign-flip AUC of the cosine score on 5G-NIDD 0.995 for every G; $G = 10$ loses 0.8–6.8 accuracy points and $G = 160$ gains 0.6–1.2 over $G = 40$; 5G-NIDD upload 59.0 KB at $G = 40$, 236 KB at $G = 160$
Bins	32, 64, 128, 256 (5 seeds)	cosine AUC 0.98–1.00 on sign-flip and permutation; accuracy within 0.2 points
Aggregator	trimmed mean vs median	pooled sign-flip AUC 0.987 vs 0.995; accuracy within 0.3 points, except sign-flip at $\rho = 0.4$ on ToN-IoT (0.864 vs 0.897)
Interactions	federated EBM, 10 shared pair terms	accuracy -0.5 / $+0.5$ / $+0.2$ points (5G-NIDD / ToN-IoT / MQTTset); pair upload 576 KB, ten times the main-effect array on 5G-NIDD

S1 Power and ablations

Figure S1 tracks recall across malicious fractions for the three attacks of the confirmation split.

Table S1 isolates the null. Every null estimated from the participants' own scores exceeded the nominal level in some exploration cell, and rescaling the simulation null by the benign core re-imported the problem. The empirical-tail simulation null made no false accusation but had almost no power at $\rho = 0.2$. Its sign-flip recall jumped only at $\rho \geq 0.3$ (0.75–1.00), when three or more attackers share the minimum attainable p-value. Table S2 shows that the other components matter little for detection. Grid size, bins and aggregator moved the cosine AUC on sign-flip and permutation by at most 0.01, and interactions did not justify their upload size.

S2 Score-level comparison, heterogeneity sweep and larger federations

Table S3 and Fig. S3 compare rankings independently of any decision rule. On sign-flip and permutation the cosine score reached AUC 0.93–1.00 at every ρ on MQTTset, 5G-NIDD and ToN-IoT, on par with FLAME. It ranked above Multi-Krum and sum-L_2 in every cell and above the per-feature statistic, whose AUC was 0.49–0.88. SHAP-on-upload stayed at 0.37–0.62 on the model-level attacks, near or below chance, as

Table S3 Score AUC on the same uploads. MQTTset, 5G-NIDD and ToN-IoT: design set and exploration split ($\alpha = 0.5$, 10 seeds), mean over seeds at $\rho = 0.1$ / $\rho = 0.4$. CICIoT2023: confirmation split, pooled over $\alpha \in \{0.5, 1, 100\}$ and $\rho \in \{0.1, \ldots, 0.4\}$; feature inversion was not run there. Ours is the cosine score, numerically identical to cosine-to-median.

Dataset	Attack	Ours	Sum-L_2	SHAP-on-upload	SHERPA-on-upload	Multi-Krum	FLAME
MQTTset	sign-flip	1.00 / 0.98	0.87 / 0.78	0.57 / 0.47	0.95 / 0.68	0.96 / 0.82	0.97 / 0.96
	permutation	0.99 / 0.93	0.83 / 0.75	0.62 / 0.53	0.95 / 0.69	0.88 / 0.78	0.95 / 0.91
	label-flip	0.85 / 0.86	0.56 / 0.05	0.64 / 0.65	0.91 / 0.64	0.43 / 0.01	0.87 / 0.84
	feature-inv.	0.78 / 0.76	0.70 / 0.66	0.63 / 0.59	0.82 / 0.64	0.69 / 0.63	0.72 / 0.74
5G-NIDD	sign-flip	0.99 / 1.00	0.81 / 0.71	0.50 / 0.43	0.93 / 0.85	0.95 / 0.82	0.99 / 1.00
	permutation	0.99 / 0.99	0.79 / 0.69	0.48 / 0.43	0.87 / 0.71	0.95 / 0.79	0.99 / 1.00
	label-flip	1.00 / 1.00	0.95 / 0.11	0.85 / 0.68	0.98 / 0.93	0.89 / 0.00	1.00 / 0.99
	feature-inv.	0.65 / 0.67	0.61 / 0.55	0.58 / 0.48	0.54 / 0.60	0.59 / 0.57	0.65 / 0.68
ToN-IoT	sign-flip	1.00 / 1.00	0.95 / 0.83	0.50 / 0.38	0.93 / 0.73	0.99 / 0.85	1.00 / 1.00
	permutation	1.00 / 0.99	0.95 / 0.80	0.51 / 0.41	0.83 / 0.69	0.98 / 0.83	1.00 / 1.00
	label-flip	0.99 / 0.95	0.61 / 0.02	0.76 / 0.56	0.93 / 0.70	0.65 / 0.00	1.00 / 0.97
	feature-inv.	0.65 / 0.57	0.52 / 0.55	0.60 / 0.54	0.51 / 0.56	0.50 / 0.52	0.72 / 0.59
CICIoT2023	sign-flip	1.00	0.90	0.53	0.96	0.96	1.00
	permutation	0.99	0.86	0.53	0.82	0.94	0.99
	label-flip	0.96	0.60	0.82	0.95	0.47	0.93

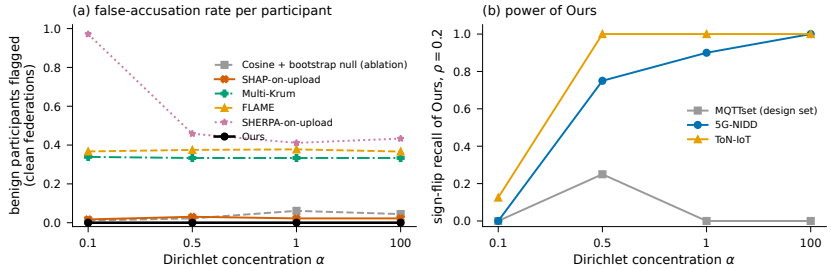

Fig. S2 Heterogeneity sweep (design set and exploration split; Ours: 10 seeds, 4–5 at $\alpha = 0.1$; comparators: 5–10 seeds). (a) Share of benign participants accused in clean federations by each detector's native rule, pooled over datasets; (b) sign-flip recall of Ours at $\rho = 0.2$.

the proposition of article Section 5 predicts. Because SHERPA-on-upload computes signed per-sample attributions, its underlying signal remains sensitive to both attacks, and its weakness here lies in its mean-threshold decision rule. In paired bootstrap tests with BH correction, the cosine score was significantly higher than sum-L_2 in 38 of 48 (dataset, attack, ρ) cells. It was higher than SHAP-on-upload in 40 cells and than Multi-Krum in 28. The cosine score was significantly lower in only 3 of 432 comparisons, all against a TreeSHAP importance audit of local LightGBM models under ToN-IoT feature inversion at $\rho \geq 0.2$. Under label flipping at large ρ, sum-L_2 inverted, whereas the scale-invariant cosine distance did not.

Figure S2 sweeps the Dirichlet concentration parameter α from strongly non-IID to nearly IID partitions. The native rules of FLAME and Multi-Krum accused 32–42% and 33–35% of benign participants in clean federations at every α from 0.1 to 100 and on every dataset. SHERPA-on-upload accused 37–48% at $\alpha \geq 0.5$ and 91–100% at $\alpha = 0.1$. In this sweep the audit accused no benign participant. At $\alpha = 0.1$ some participants and virtual participants held a single class. In this sweep sign-flip recall at $\rho = 0.2$ was 0.00–0.12, with no false accusation in 4–5 clean federations per dataset.

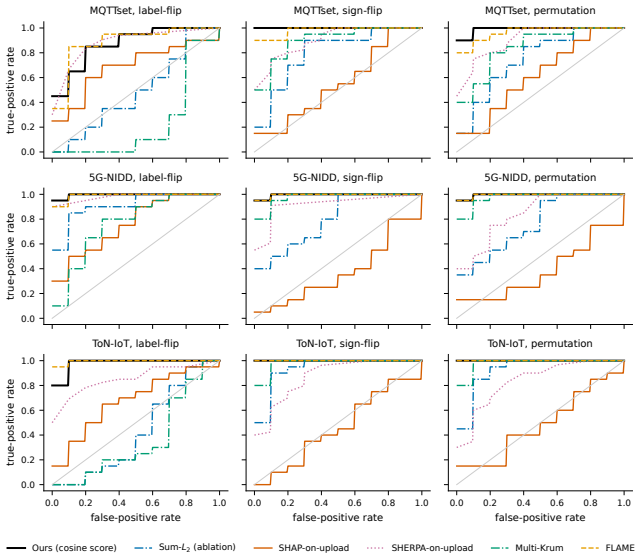

Fig. S3 Receiver operating characteristic curves at $\rho = 0.2$ (vertical average over 10 seeds; design set and exploration split, $\alpha = 0.5$). Ours is the cosine score.

Table S4 The audit at $K = 12$, 24 and 48 ($\alpha = 0.5$, 10 seeds; MQTTset, 5G-NIDD and ToN-IoT). The $K = 12$ rows are stored outputs of the same method (5G-NIDD and ToN-IoT: confirmation seeds 10–19; MQTTset: seeds 0–9) from a different seed set than the $K = 24$ and 48 rows. The p-values of the $K = 24$ and 48 rows were computed in the exploration step and analyzed afterwards.

Dataset	K	Clean feds with false acc.	Median p	SF recall ρ=0.1 / 0.2 / 0.3	PM 0.2	LF 0.2	AUC SF 0.2
MQTTset	12	0/10	0.61	0.20 / 0.25 / 0.30	0.00	–	1.00
	24	0/10	0.57	0.00 / 0.00 / 0.00	0.00	0.00	0.99
	48	0/10	0.58	0.00 / 0.01 / 0.00	0.01	0.00	0.98
5G-NIDD	12	0/10	0.55	0.90 / 0.90 / 0.97	0.55	0.65	1.00
	24	0/10	0.59	0.50 / 0.88 / 0.94	0.48	0.52	1.00
	48	0/10	0.54	0.12 / 0.12 / 0.26	0.00	0.13	0.99
ToN-IoT	12	0/10	0.69	1.00 / 1.00 / 1.00	1.00	0.50	1.00
	24	0/10	0.61	1.00 / 1.00 / 1.00	0.98	0.32	1.00
	48	0/10	0.63	1.00 / 1.00 / 1.00	1.00	0.26	1.00

The sweep excluded 5G-NIDD seed 3, which had only 44 usable virtual participants. The $\alpha = 0.1$ row of article Table 4 and Table S25 scores five federations per dataset, and that seed with its 44 virtual participants is among them. Its sign-flip recall was 0.00–0.10 per dataset.

With $K = 24$ and 48 participants no clean federation had a false accusation (0/60, Table S4). Power fell with K on 5G-NIDD, to 0.12 at $K = 48$, and MQTTset fell to near zero. The score still ranked attackers (sign-flip AUC 0.98–1.00 at $\rho = 0.2$), and the cosine decomposition named the attacked feature in every model-level attacked upload. Multi-Krum degraded on label flipping at $K \geq 24$ (AUC 0.02–0.08).

Table S5 Attribution hit rate: share of attacked uploads whose top-attributed feature is the attacked one, by cosine decomposition (cos) and standardized attribution (z); benign control: share of benign participants whose top-attributed feature is their own most influential one. Left: design set and exploration split ($\alpha = 0.5$, 120 attacked uploads per cell); right: confirmation split ($\rho = 0.1$–0.4, 360 uploads per dataset and attack), including the accused attackers. "–": not part of that split.

	Design set and exploration split					Confirmation split			
Dataset (F)	sign-flip cos / z	permutation cos / z	feature-inv. cos / z	benign control cos / z	chance $1/F$	sign-flip cos / z	permutation cos / z	accused SF / PM cos	benign control cos / z
MQTTset (16)	1.00 / 0.72	1.00 / 0.71	0.67 / 0.50	0.61 / 0.24	0.063	–	–	–	–
5G-NIDD (41)	1.00 / 0.55	1.00 / 0.55	0.33 / 0.29	0.48 / 0.03	0.024	1.000 / 0.717	1.000 / 0.692	1.000 / 1.000	0.386 / 0.017
ToN-IoT (15)	1.00 / 0.58	1.00 / 0.52	0.10 / 0.42	0.66 / 0.11	0.067	1.000 / 0.625	1.000 / 0.589	1.000 / 1.000	0.661 / 0.161
CICIoT2023 (39)	–	–	–	–	0.026	0.989 / 0.442	0.989 / 0.297	1.000 / 1.000	0.972 / 0.019

Table S6 Benign holder of a novel class under the audit (unrestricted score; $\alpha = 0.5$, $K = 12$; 15 federations per class: 5 seeds $\times$ $\rho \in \{0, 0.1, 0.2\}$): the holder keeps its rows of the class and every other participant drops them. Columns: holder accused; holder's median p-value; other benign participants accused; sign-flip recall and score AUC of the concurrent attackers ($\rho > 0$).

Dataset	Novel class	Holder accused	Holder median p	Other benign accused	SF recall	AUC
MQTTset	`bruteforce`	0/15	0.61	0/150	0/15	1.00
	`malformed`	0/15	0.44	0/150	0/15	1.00
5G-NIDD	`HTTPFlood`	0/15	0.09	1/150	10/15	0.99
	`SlowrateDoS`	0/15	0.32	3/150	13/15	0.99
ToN-IoT	`password`	0/15	0.55	0/150	15/15	1.00
	`ransomware`	3/15	0.16	0/150	15/15	1.00
CICIoT2023	`BruteForce`	0/15	0.67	0/150	13/15	1.00
	Web	0/15	0.57	0/150	14/15	1.00

S3 Attribution over all malicious fractions

Over all malicious fractions the standardized attribution hit 0.297–0.717 of the attacked uploads on the confirmation split, against a control of 0.017–0.161 (Table S5). Its hit rate is higher at $\rho \leq 0.2$ (article Table 5), which is consistent with more attackers distorting the robust z-score across participants. These rates are over all attacked uploads, most of them unaccused, and Table S19 gives the rates conditional on accusation. For feature inversion, attribution was above chance on MQTTset and 5G-NIDD and weak for the cosine decomposition on ToN-IoT. For accused label flippers, the named class column gained rows from the flip in 0.368, 1.000 and 0.667 of the cases on 5G-NIDD, ToN-IoT and CICIoT2023. These shares rest on 95, 23 and 3 accused attackers.

S4 Novelty triage

The holder of a single-holder class was accused in 3 of 120 federations (0.025 [0.005, 0.071]) (Table S6). All three are one federation, ToN-IoT `ransomware` seed 1 ($p \approx 0.004$), at the three values of ρ. The rate was 0 in the other 21 (class, ρ) cells, including 5G-NIDD `HTTPFlood`, whose holder the two-pass bootstrap null had accused in 20–40% of federations. Other benign participants were accused in 4 of 1,200 cases, none in a clean federation. The novelty signal is a deterministic property, not a detection rate, because in every federation the holder was the only participant with non-zero

Table S7 Computation ($\alpha = 0.5$). Participant rows: one idle thread, mean of the smallest, median and largest partition, multilayer perceptron (MLP) memory including its framework; server rows: one to three federations per dataset. CPU-min: minutes of central processing unit time; [a]estimated from stored fit times scaled by idle timings (one 5G-NIDD audit measured: 7.8 CPU-min); [b]recorded under load.

	MQTTset	5G-NIDD	ToN-IoT	CICIoT2023
MLP rounds to best validation	5–89	25–106	52–150	28–87
EBM fit per participant (s)	54.4	104.3	66.7	46.8
MLP local round per participant (s)	0.20	0.15	0.11	0.074[b]
Peak memory, EBM / MLP (MB)	182 / 398	198 / 410	177 / 391	194 / –
Score, p-values and BH, $K = 12$ (ms)	2.0	9.2	3.0	7.6
Simulation-null audit, one thread (CPU-min)[a]	7.4	11.5	6.1	6.3
Simulation-null audit, 16 workers, full load (wall s)	239	607	291	234
Bootstrap-null audit, $K = 12$ (s)	0.19	1.61	0.31	1.42

columns on the unsupported class. The server's pool contained rows of the single-holder class, and the triage was not tested against a malicious participant that edits the same columns.

S5 Cost

On the server, the decision step took under 10 ms, and generating the simulation null required fitting 48 virtual participants (Table S7). All timings come from one Intel Core i7-13620H processor under Windows 11, with 6 performance and 4 efficiency cores, 16 threads and a 2.4 GHz performance-core base clock. A single 5G-NIDD audit required 7.8 CPU-minutes on an idle thread, and estimated runtimes across 12 federations spanned 5.4–16.3 CPU-minutes. This runtime is roughly 400 times the CPU time of the fit-free bootstrap-null audit. A virtual fit took 9.3 s on average (range 2.0–65 s), less than a participant fit. With 16 parallel workers an audit took 2.7–14 minutes (mean 5.7) on a fully loaded machine, and each fit added at most 7 MB of peak memory. On the participant's side, an EBM fit takes 3–55 times the cumulative local training of the multilayer perceptron (MLP) up to its best validation round on MQTTset, 5G-NIDD and ToN-IoT. It uses roughly half the MLP's peak memory. On matching seeds, federating under non-IID data lost 3.4, 2.4 and 5.4 accuracy points relative to a centralized EBM on MQTTset, 5G-NIDD and ToN-IoT.

S6 Attacks

The four full-knowledge attacks of article Section 7.4 have access to every benign upload and act on the flattened arrays. ALIE uploads the coordinate-wise benign mean minus $z_{\max}$ times the coordinate-wise standard deviation, with $z_{\max}$ from Baruch et al. [1] for K and the number of attackers. Min-Max adds γ times the negative standard deviation to the benign mean, with γ found by bisection. The chosen γ is the largest whose maximum distance to any benign upload does not exceed the largest benign pairwise distance. Min-Sum is the same with the constraint on the sum of squared distances [6]. The trimmed mean attack [3], which its authors also use against the median, pushes each coordinate beyond the benign extreme opposite to the benign mean direction. It scales that extreme by a factor drawn uniformly in $[1, 2]$ when the

Table S8 Standard attacks on the confirmation federations ($\alpha = 0.5$, 10 seeds): the 10 of 72 (dataset, aggregator, attack, ρ) cells with at least 1 pp of accuracy damage without audit, and the fraction recovered by excluding the participants each rule accuses. Exclusion cost in clean federations: Ours 0.00 pp, FLAME up to 1.23 pp, Multi-Krum up to 0.77 pp.

						Recovered fraction		
Dataset	Aggregator	Attack	ρ	Damage (pp)	Recall of Ours	Ours	FLAME	Multi-Krum
ToN-IoT	median	label-flip	0.4	1.06	0.20	0.06	0.53	-2.31
ToN-IoT	mean	sign-flip	0.3	1.44	1.00	0.52	0.68	0.41
ToN-IoT	mean	sign-flip	0.4	5.33	1.00	0.88	0.89	0.44
ToN-IoT	mean	permutation	0.4	2.25	1.00	0.71	0.73	0.22
CICIoT2023	median	sign-flip	0.4	2.40	0.84	0.91	0.87	-0.12
CICIoT2023	median	permutation	0.4	1.11	0.36	0.28	0.75	-0.66
CICIoT2023	mean	sign-flip	0.3	3.93	0.95	0.93	0.95	0.88
CICIoT2023	mean	sign-flip	0.4	9.81	0.84	0.94	0.95	0.65
CICIoT2023	mean	permutation	0.3	1.08	0.60	0.63	0.81	0.49
CICIoT2023	mean	permutation	0.4	2.04	0.36	0.38	0.66	0.23

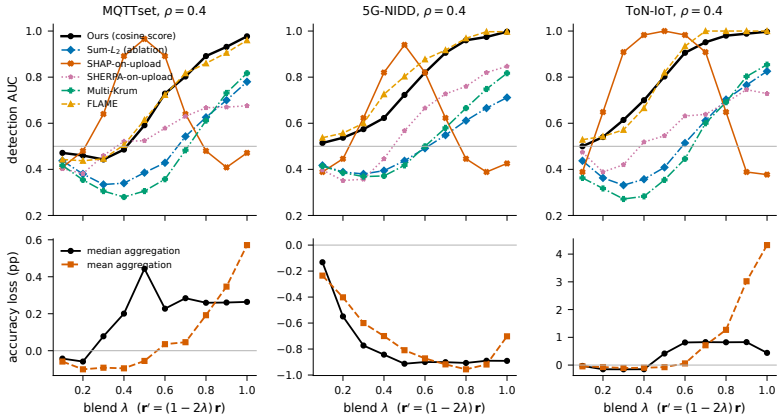

Fig. S4 Stealth sweep at $\rho = 0.4$ (design set and exploration split, $\alpha = 0.5$, 10 seeds): the attacker uploads $(1-2\lambda)\mathbf{r}$ for its most influential feature. Top: score AUC. Bottom: accuracy loss of the shared detector (percentage points) under median and mean aggregation; negative values are gains.

extreme has the same sign as the push and divides it by the factor otherwise. All attackers upload the same crafted vector, except under the trimmed mean attack, where each attacker draws its factors independently.

Table S8 lists the standard-attack cells with at least 1 point of damage. Recovery was lowest for ToN-IoT label flipping under the median, where the audit's recall is lowest. FLAME's exclusions recovered as much or more in 9 of the 10 cells, the exception being CICIoT2023 sign-flip under the median at $\rho = 0.4$ (0.91 against 0.87).

No audit we evaluated detected the targeted evasion-window attack. A region-level scan, the maximum robust z-score over (feature, class, window) cells with settings selected on the design set, had AUC 0.49–0.52 on the exploration split at $\rho \in \{0.1, 0.2\}$. This AUC fell below that of the cosine score (0.59–0.73) and FLAME (0.56–0.79). Under the simulation nulls the cosine audit accused none of these attackers at $\rho = 0.2$. Under median aggregation the recall of the target class inside the window fell by at most 0.4 points on 5G-NIDD and ToN-IoT (0.9 on MQTTset, interval including 0).

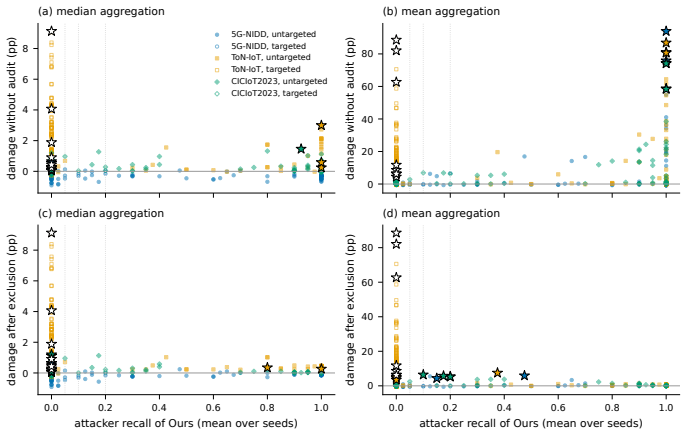

Fig. S5 Grid oracle over the blend-and-boost family ($\alpha = 0.5$, $K = 12$): damage against attacker recall of Ours per dataset, goal, ρ and configuration (mean over 10 seeds). Rows: without audit and after exclusion; filled: untargeted accuracy loss; open: targeted $c^\star$ recall loss; stars: grid-oracle picks; dotted: detection budgets 0.05, 0.1 and 0.2.

Under mean aggregation it fell by up to 3.4 points. In the stealth sweep of Fig. S4, the cosine score had AUC 0.47–0.51 at $\lambda = 0.1$, 0.59–0.80 at $\lambda = 0.5$ and 0.98–1.00 at $\lambda = 1$. SHAP-on-upload behaved in the opposite way, with 0.94–1.00 at $\lambda = 0.5$ and 0.38–0.47 at $\lambda = 1$, where only the sign changes. The accuracy loss was at most 1 point for every λ with median aggregation and at most 4.3 points with mean aggregation.

Table S9 gives the grid oracle of article Section 7.4 per cell, and Fig. S5 plots its damage against attacker recall. It knows the detection rate of every configuration of the blend-and-boost family, so its damage is an upper bound within this family only. In 2 of the 12 cells at $\rho \leq 0.2$, both untargeted on ToN-IoT, the oracle configuration was accused (recall 1.00 and 0.80). Exclusion left accuracy damage of 0.27 and 0.35 points there. At $\rho = 0.3$ three cells exceeded 1 point of accuracy or 5 points of recall after exclusion, all with attacker recall 0.00. On ToN-IoT a small untargeted edit spread over all features ($d = F$, $\lambda = 0.4$, $\gamma = 1$) removed 1.11 points of accuracy. On ToN-IoT a boosted targeted edit ($d = F$, $\lambda = 0.6$) removed 9.13 points of recall of `password`, while overall accuracy fell from 90.0% to 88.8%. On CICIoT2023 a boosted untargeted edit of one feature ($d = 1$, $\lambda = 0.5$) removed 1.19 points of accuracy. Under mean aggregation, exclusion reduced the untargeted damage of boosted full sign-flips to 3.4–7.5 points. The targeted class column carried only 0.06–0.18 of the consensus energy $\|\bar{\mathbf{S}}\|_2^2$, so negating it moved $1 - \cos$ little, and its smallest p-value was never below 0.013.

For each class c that a participant holds, the per-class score tests the cosine distance of its class column to the consensus column. The null for c uses the virtual participants that hold c, with the same log-t tail, when at least 8 of them do. A participant's p-value is the Bonferroni minimum over its classes, and BH at $q = 0.1$ runs across participants. Its false accusations in clean federations (Table S11) occurred in 1 of 30 on 5G-NIDD, 2 of 30 on ToN-IoT and 1 of 30 on CICIoT2023. All its grid-oracle configurations kept attacker recall at 0.00 on ToN-IoT.

Table S9 Grid oracle over the blend-and-boost family per cell ($\alpha = 0.5$, 10 seeds): largest mean damage over the 54 configurations of each goal without audit ("none") and after excluding the accused ("after"), and the configuration that maximizes damage after exclusion with its attacker recall. Damage in percentage points of accuracy (U) and of $c^\star$ recall (T); d: number of edited features; the audit uses the median consensus.

			Median aggregation			Mean aggregation		
Dataset	Goal	ρ	none	after	oracle configuration; recall	none	after	oracle configuration; recall
5G-NIDD	U	0.1	0.20	0.20	d=1, λ=0, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	75.8	5.4	d=1, λ=0.8, γ=$K/\lvert\mathcal{M}\rvert$; 0.20
5G-NIDD	U	0.2	0.55	0.55	d=1, λ=0, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	81.0	4.4	d=1, λ=0.8, γ=$K/\lvert\mathcal{M}\rvert$; 0.15
5G-NIDD	U	0.3	0.64	0.64	d=1, λ=0, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	94.0	5.9	d=1, λ=1, γ=$K/\lvert\mathcal{M}\rvert$; 0.47
5G-NIDD	T	0.1	0.06	0.06	d=F, λ=0.4, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	6.4	6.4	d=F, λ=1, γ=$K/\lvert\mathcal{M}\rvert$; 0.00
5G-NIDD	T	0.2	0.04	0.04	d=F, λ=0.3, γ=1; 0.00	4.4	4.4	d=F, λ=1, γ=$K/\lvert\mathcal{M}\rvert$; 0.00
5G-NIDD	T	0.3	0.06	0.06	d=F, λ=0.4, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	6.2	6.2	d=F, λ=1, γ=$K/\lvert\mathcal{M}\rvert$; 0.00
ToN-IoT	U	0.1	0.24	0.27	d=1, λ=0.8, γ=1; 1.00	74.3	5.1	d=1, λ=0, γ=$K/\lvert\mathcal{M}\rvert$; 0.00
ToN-IoT	U	0.2	0.58	0.35	d=F, λ=0.5, γ=1; 0.80	80.6	3.4	d=1, λ=0, γ=$K/\lvert\mathcal{M}\rvert$; 0.00
ToN-IoT	U	0.3	3.00	1.11	d=F, λ=0.4, γ=1; 0.00	86.8	7.5	d=1, λ=0.8, γ=$K/\lvert\mathcal{M}\rvert$; 0.38
ToN-IoT	T	0.1	1.89	1.89	d=F, λ=0.6, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	62.7	62.7	d=F, λ=1, γ=$K/\lvert\mathcal{M}\rvert$; 0.00
ToN-IoT	T	0.2	4.07	4.07	d=F, λ=0.8, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	82.1	82.1	d=F, λ=1, γ=$K/\lvert\mathcal{M}\rvert$; 0.00
ToN-IoT	T	0.3	9.13	9.13	d=F, λ=0.6, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	88.6	88.6	d=F, λ=1, γ=$K/\lvert\mathcal{M}\rvert$; 0.00
CICIoT2023	U	0.1	0.09	0.09	d=1, λ=0.5, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	58.7	6.4	d=1, λ=0.6, γ=$K/\lvert\mathcal{M}\rvert$; 0.10
CICIoT2023	U	0.2	0.28	0.28	d=1, λ=0.5, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	58.4	5.2	d=1, λ=0.6, γ=$K/\lvert\mathcal{M}\rvert$; 0.20
CICIoT2023	U	0.3	1.46	1.19	d=1, λ=0.5, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	74.1	5.7	d=1, λ=0.6, γ=$K/\lvert\mathcal{M}\rvert$; 0.17
CICIoT2023	T	0.1	0.23	0.23	d=1, λ=0.5, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	8.6	8.6	d=F, λ=1, γ=$K/\lvert\mathcal{M}\rvert$; 0.00
CICIoT2023	T	0.2	0.43	0.43	d=1, λ=0.5, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	6.2	6.2	d=F, λ=1, γ=$K/\lvert\mathcal{M}\rvert$; 0.00
CICIoT2023	T	0.3	0.90	0.90	d=1, λ=0.5, γ=$K/\lvert\mathcal{M}\rvert$; 0.00	11.7	11.7	d=F, λ=1, γ=$K/\lvert\mathcal{M}\rvert$; 0.00

Table S10 Full-knowledge attacks under median aggregation (30 confirmation federations per dataset: 3 α × 10 seeds), at $\rho = 0.1$ / 0.2 / 0.3 / 0.4. Damage in percentage points against the clean federation, without audit and after exclusion of the accused ("none, after"); worst: recall of the class with the lowest clean recall. FA: attacked federations in which a benign participant was accused, of 30.

Attack	Dataset	Recall	Accuracy damage (none, after)	Worst-class recall damage (none, after)	FA
ALIE	5G-NIDD	0.00 / 0.00 / 0.00 / 0.00	-0.07, -0.07 / -0.16, -0.16 / -0.57, -0.57 / -0.74, -0.74	-2.52, -2.52 / -2.31, -2.31 / 1.22, 1.22 / 9.13, 9.13	0 / 0 / 0 / 0
ALIE	ToN-IoT	0.00 / 0.00 / 0.00 / 0.00	0.09, 0.09 / 0.10, 0.10 / 0.16, 0.16 / 0.44, 0.43	-0.68, -0.68 / 0.97, 0.97 / 9.15, 9.15 / 16.73, 16.73	0 / 0 / 0 / 1
ALIE	CICIoT2023	0.00 / 0.00 / 0.00 / 0.00	0.00, 0.00 / 0.02, 0.02 / 0.13, 0.13 / 0.33, 0.33	-0.62, -0.62 / -0.86, -0.86 / -0.08, -0.08 / -0.17, -0.17	0 / 0 / 0 / 0
Min-Max	5G-NIDD	0.00 / 0.00 / 0.00 / 0.00	-0.14, -0.14 / -0.29, -0.29 / -0.58, -0.58 / -0.84, -0.84	3.18, 3.18 / 1.01, 1.01 / 16.53, 16.53 / 18.12, 18.12	0 / 0 / 0 / 0
Min-Max	ToN-IoT	0.00 / 0.00 / 0.00 / 0.00	0.06, 0.06 / 0.03, 0.03 / 0.29, 0.29 / 0.80, 0.80	2.48, 2.48 / 5.33, 5.33 / 19.47, 19.47 / 24.15, 24.15	0 / 0 / 1 / 0
Min-Max	CICIoT2023	0.00 / 0.00 / 0.00 / 0.00	-0.03, -0.03 / 0.01, 0.01 / 0.30, 0.30 / 0.59, 0.59	-0.05, -0.05 / 0.21, 0.21 / -0.15, -0.15 / -0.48, -0.48	0 / 0 / 0 / 0
Min-Sum	5G-NIDD	0.00 / 0.00 / 0.00 / 0.00	-0.14, -0.14 / -0.29, -0.29 / -0.61, -0.61 / -0.80, -0.80	3.18, 3.18 / 1.01, 1.01 / 15.99, 15.99 / 16.18, 16.18	0 / 0 / 0 / 0
Min-Sum	ToN-IoT	0.00 / 0.00 / 0.00 / 0.00	0.06, 0.06 / 0.03, 0.03 / 0.26, 0.26 / 0.65, 0.65	2.48, 2.48 / 5.33, 5.33 / 18.68, 18.68 / 23.08, 23.08	0 / 0 / 1 / 0
Min-Sum	CICIoT2023	0.00 / 0.00 / 0.00 / 0.00	-0.03, -0.03 / 0.01, 0.01 / 0.29, 0.29 / 0.50, 0.50	-0.05, -0.05 / 0.21, 0.21 / -0.27, -0.27 / -0.41, -0.41	0 / 0 / 0 / 0
Trimmed mean attack	5G-NIDD	0.40 / 0.53 / 0.53 / 0.23	-0.16, -0.07 / -0.40, -0.15 / -0.72, -0.45 / 0.92, 0.42	4.69, -2.77 / 6.40, -0.92 / 23.42, 2.38 / 41.81, 27.41	0 / 0 / 0 / 0
Trimmed mean attack	ToN-IoT	0.97 / 0.97 / 0.77 / 0.00	0.01, 0.09 / -0.05, 0.00 / 1.33, 0.18 / 11.52, 11.52	7.61, 2.58 / 14.78, 1.77 / 38.36, 10.02 / 59.43, 59.43	1 / 1 / 1 / 0
Trimmed mean attack	CICIoT2023	0.53 / 0.67 / 0.67 / 0.00	-0.06, -0.04 / -0.03, -0.07 / 0.98, 0.00 / 6.47, 6.47	1.16, 0.59 / 2.58, 0.87 / 3.48, 0.66 / 3.85, 3.85	0 / 0 / 0 / 0

Table S11 Per-class score against the cosine score on the confirmation federations. Clean: clean federations with a false accusation, with exact 95% interval; LF recall: pooled label-flip recall at $\rho = 0.1$ / 0.2 / 0.3 / 0.4. Targeted damage: grid-oracle damage to the recall of $c^\star$ after exclusion on ToN-IoT at $\rho = 0.1$ / 0.2 / 0.3 (pp), with the grid oracle re-selected per score.

Score	Clean	LF recall	Targeted damage
Cosine (method)	0/90; clusters 0/30 [0.000, 0.116]	0.16 / 0.27 / 0.35 / 0.30	1.89 / 4.07 / 9.13
Per-class	4/90 [0.012, 0.110]	0.48 / 0.52 / 0.56 / 0.65	1.44 / 3.53 / 9.07

S7 Adaptation of ByMI at 48 participants

The adapted ByMI follows Section 2.2 of its article [4], with the response-curve upload of each half in place of a gradient (Table S12). Every half upload was normalized to unit length with weighting matrix $\mathbf{\Omega} = \mathbf{I}$, so W_k forms the cross-fitted counterpart of the cosine score. With the full-upload centre, W_k need not be centred at zero even for

Table S12 The adaptation of ByMI and the audit on the $K = 48$ federations ($\alpha = 0.5$; MQTTset, 5G-NIDD and ToN-IoT, 10 seeds each). Any-FA: federations with a benign participant accused, of 30 (of 9 for the IID control), with exact 95% interval (%). Benign FA: share of benign participants accused (%; counts for the IID control and the lower block); FDP: mean false-discovery proportion (%); recall in %. IID control: nine federations with a random equal stratified partition, every participant benign. Lower block: other statistics and centres on clean federations.

Method	Condition	Any-FA	Benign FA	Recall	FDP
Audit	clean	0/30 [0.0, 11.6]	0.0	–	–
	sign-flip, $\rho = 0.1$ / 0.2	1/30 / 4/30	0.1 / 0.7	37.3 / 37.7	0.6 / 2.1
	permutation, $\rho = 0.1$ / 0.2	1/30 / 3/30	0.1 / 0.6	30.7 / 33.7	0.7 / 1.8
Adapted ByMI, centre: median of full uploads	clean	30/30 [88.4, 100.0]	95.3	–	–
	sign-flip, $\rho = 0.1$ / 0.2	30/30 / 30/30	96.3 / 96.9	100.0 / 100.0	89.2 / 78.7
	permutation, $\rho = 0.1$ / 0.2	30/30 / 30/30	96.3 / 96.8	100.0 / 100.0	89.2 / 78.7
Adapted ByMI, centre: median of first halves	clean	30/30 [88.4, 100.0]	96.3	–	–
	sign-flip, $\rho = 0.1$ / 0.2	30/30 / 30/30	96.8 / 97.5	100.0 / 100.0	89.3 / 78.9
	permutation, $\rho = 0.1$ / 0.2	30/30 / 30/30	96.8 / 97.4	100.0 / 100.0	89.3 / 78.8
Adapted ByMI, full-upload centre	IID control	8/9 [51.8, 99.7]	211/432	–	–
Adapted ByMI, first-half centre	IID control	3/9 [7.5, 70.1]	43/432	–	–
EBM uploads, no normalization, full-upload centre	IID control	6/9 [29.9, 92.5]	275/432	–	–
EBM uploads, no normalization, first-half centre	IID control	0/9 [0.0, 33.6]	0/432	–	–
Linear class-mean statistic, full-upload centre	IID control	0/9 [0.0, 33.6]	0/432	–	–
Linear class-mean statistic, first-half centre	IID control	2/9 [2.8, 60.0]	20/432	–	–
Linear class-mean statistic, full-upload centre	$K = 48$, clean	12/30 [22.7, 59.4]	279/1438	–	–
Linear class-mean statistic, first-half centre	$K = 48$, clean	13/30 [25.5, 62.6]	307/1438	–	–
Linear, class blocks in both halves, either centre	IID control; $K = 48$	0/9; 0/30	0	–	–

IID benign participants. The centre is fitted on twice the data of a half, and the median has its own error in 3,840 to 14,760 coordinates. Label skew adds a participant-specific positive component. The IID control separates these effects, and the false accusations of the adapted ByMI persisted there with both centres. At $q = 0.1$ the threshold of ByMI rejects zero or at least 10 participants, so at $\rho = 0.1$ (5 attackers of 48) any rejection set holds at least 5 benign participants.

The lower block of Table S12 varies the statistic and the centre on clean federations. A linear statistic of class-conditional feature means has independent and unbiased halves under IID data. It rejected in none of the 9 IID federations with the full-upload centre and in 2 of 9 with the first-half centre. Zero-filled columns of classes absent from one half may produce such rejections. A variant restricted to class blocks present in both halves of every participant rejected in none of the 39 federations. The adaptation thus failed even with a linear statistic, and these runs make no statement about ByMI in its own setting.

S8 Details of the calibration evidence

In the in-sample score the virtual participants scored higher than the benign ones in the median federation of every cell (Table S13). With the leave-one-out score the median gap turned positive at $\alpha = 0.5$ on 5G-NIDD and ToN-IoT, yet no clean federation had a false accusation. No benign p-value fell below q/K, and the smallest one was 1.80 times that threshold (Table S14). In the leave-one-virtual-out check each virtual participant was scored against the other 47 with the same log-t tail. The share of these p-values below 0.01 stayed at most 0.0063 in every cell.

Table S15 splits the altered conditions by dataset. In the out-of-capture pool test the 5G-NIDD station of the participants held 55.7% benign and 24.2% `UDPFlood` rows, against 14.4% and 58.5% at the station of the pool. Both stations come from the same

Table S13 Benign-minus-virtual log-score gap per clean federation, the median of $\log s$ over benign real participants minus the median of $\log s_v$ over virtual participants. Entries: median over the 20 clean federations of each cell (confirmation and unseen seeds) [minimum, maximum; federations with a gap of at least zero]. Columns: the in-sample score of the method and the leave-one-out (LOO) score. Negative values mean that the virtual participants score higher than the benign ones.

Dataset	α	In-sample	LOO
5G-NIDD	0.5	−0.132 [−0.401, 0.039; 3]	0.072 [−0.252, 0.390; 14]
	1	−0.858 [−1.098, −0.260; 0]	−0.657 [−0.906, −0.034; 0]
	100	−1.417 [−1.919, −1.077; 0]	−1.242 [−1.738, −0.927; 0]
ToN-IoT	0.5	−0.168 [−0.284, 0.046; 2]	0.018 [−0.121, 0.221; 11]
	1	−0.360 [−0.642, −0.007; 0]	−0.188 [−0.450, 0.163; 1]
	100	−0.797 [−1.122, −0.574; 0]	−0.590 [−0.980, −0.393; 0]
CICIoT2023	0.5	−0.156 [−0.576, 0.086; 3]	−0.011 [−0.410, 0.256; 10]
	1	−0.782 [−1.154, −0.570; 0]	−0.618 [−0.970, −0.415; 0]
	100	−1.468 [−1.645, −1.196; 0]	−1.259 [−1.469, −0.965; 0]

Table S14 Benign p-values in the 20 clean federations of each cell (confirmation and unseen seeds, 240 benign participants per cell). Counts below q/K, $2q/K$ and q; smallest benign p-value per federation as a ratio to the first BH threshold q/K (minimum, 5% quantile and median over federations). LOVO: share of leave-one-virtual-out p-values below 0.01.

		Benign p below			Smallest benign p / (q/K)			
Dataset	α	q/K	$2q/K$	q	min	5%	median	LOVO
5G-NIDD	0.5	0	1	15	1.80	2.20	9.80	0.0042
	1	0	0	3	6.09	8.87	30.39	0.0021
	100	0	0	0	79.62	83.88	100.26	0.0052
ToN-IoT	0.5	0	0	11	2.12	3.38	15.53	0.0063
	1	0	0	4	6.44	6.96	26.58	0.0021
	100	0	0	0	17.92	20.95	48.28	0.0010
CICIoT2023	0.5	0	0	12	3.40	4.74	14.51	0.0010
	1	0	0	2	8.71	10.83	52.36	0.0000
	100	0	0	0	97.85	102.99	115.44	0.0000

Table S15 Clean federations with a false accusation per dataset under the conditions of Table S25, with exact 95% intervals in % that treat the α cells of a seed as independent.

Condition	5G-NIDD	ToN-IoT	CICIoT2023
Unseen seeds	0/30 [0.0, 11.6]	0/30 [0.0, 11.6]	0/30 [0.0, 11.6]
Feature shift	0/30 [0.0, 11.6]	5/30 [5.6, 34.7]	0/30 [0.0, 11.6]
Quantity skew	1/30 [0.1, 17.2]	1/30 [0.1, 17.2]	0/30 [0.0, 11.6]
Declared metadata (reference)	0/15 [0.0, 21.8]	0/15 [0.0, 21.8]	0/15 [0.0, 21.8]
Falsified declarations	0/15 [0.0, 21.8]	1/15 [0.2, 31.9]	0/15 [0.0, 21.8]
No declarations, $\hat{N} = N$	0/15 [0.0, 21.8]	1/15 [0.2, 31.9]	0/15 [0.0, 21.8]
No declarations, $\hat{N} = 2N$	1/15 [0.2, 31.9]	0/15 [0.0, 21.8]	2/15 [1.7, 40.5]

testbed and campaign, so the test changes the station and the class mix but not the network.

S9 Calibration bounds, sampler and feature shift

Table S16 gives the bounds that the design supports. The three α cells of a confirmation or unseen seed share their data draws, so the 180 federations of these sets form

Table S16 Clean federations with a false accusation by (dataset, seed) cluster and by α cell, with exact 95% intervals. A cluster counts once if any of its federations had a false accusation; clusters are independent, and so are the federations of one α cell. New seeds: one federation per seed at $\alpha = 0.5$.

Scope	5G-NIDD	ToN-IoT	CICIoT2023	Pooled
Clusters, confirmation	0/10 [0.000, 0.308]	0/10 [0.000, 0.308]	0/10 [0.000, 0.308]	0/30 [0.000, 0.116]
Clusters, unseen seeds	0/10 [0.000, 0.308]	0/10 [0.000, 0.308]	0/10 [0.000, 0.308]	0/30 [0.000, 0.116]
Clusters, new seeds	0/20 [0.000, 0.168]	0/20 [0.000, 0.168]	1/20 [0.001, 0.249]	1/60 [0.000, 0.089]
Clusters, all three sets	0/40 [0.000, 0.088]	0/40 [0.000, 0.088]	1/40 [0.001, 0.132]	1/120 [0.000, 0.046]
$\alpha = 0.5$, confirmation and unseen	0/20 [0.000, 0.168]	0/20 [0.000, 0.168]	0/20 [0.000, 0.168]	0/60 [0.000, 0.060]
$\alpha = 1$, confirmation and unseen	0/20 [0.000, 0.168]	0/20 [0.000, 0.168]	0/20 [0.000, 0.168]	0/60 [0.000, 0.060]
$\alpha = 100$, confirmation and unseen	0/20 [0.000, 0.168]	0/20 [0.000, 0.168]	0/20 [0.000, 0.168]	0/60 [0.000, 0.060]
$\alpha = 0.5$, all three sets	0/40 [0.000, 0.088]	0/40 [0.000, 0.088]	1/40 [0.001, 0.132]	1/120 [0.000, 0.046]
$\alpha = 0.2$, new partitions	0/10 [0.000, 0.308]	0/10 [0.000, 0.308]	0/10 [0.000, 0.308]	0/30 [0.000, 0.116]
$\alpha = 0.3$, new partitions	1/10 [0.003, 0.445]	0/10 [0.000, 0.308]	0/10 [0.000, 0.308]	1/30 [0.001, 0.172]

Table S17 Federations more heterogeneous than the virtual participants ($\alpha < \alpha_v$) and a sampler matched to the federation ($\alpha_v = \alpha$), 30 federations per row. Clean: federations with a false accusation, with exact 95% interval; SF, PM: recall at $\rho = 0.2$; gap: median [interquartile range] over federations. Benign p below q/K / $2q/K$ / q, with the number of benign participants; min p: smallest benign p per federation over q/K, median and minimum.

Cell	Clean	SF	PM	Gap, in-sample	Gap, LOO	Benign p below	Min p
$\alpha = 0.2$, $\alpha_v = 0.5$	0/30 [0.000, 0.116]	0.50	0.30	−0.024 [−0.081, 0.019]	0.162 [0.106, 0.238]	0 / 2 / 37 of 359	7.78, 1.91
$\alpha = 0.3$, $\alpha_v = 0.5$	1/30 [0.001, 0.172]	0.75	0.48	−0.031 [−0.190, −0.008]	0.144 [0.019, 0.228]	1 / 4 / 37 of 360	6.90, 0.04
$\alpha = 0.2$, $\alpha_v = 0.2$	0/30 [0.000, 0.116]	0.27	0.13	−0.309 [−0.386, −0.268]	−0.118 [−0.168, −0.074]	0 / 0 / 3 of 359	22.98, 4.11
$\alpha = 100$, $\alpha_v = 0.5$	0/30 [0.000, 0.116]	0.97	0.68	−1.325 [−1.447, −0.923]	−1.141 [−1.283, −0.732]	0 / 0 / 0 of 360	100.26, 21.11
$\alpha = 100$, $\alpha_v = 100$	10/30 [0.173, 0.528]	1.00	1.00	−0.050 [−0.197, 0.030]	0.139 [−0.007, 0.208]	12 / 19 / 49 of 360	4.45, 0.00

Table S18 Per-participant feature rescaling on the confirmation federations: three random features of every real participant multiplied by factors drawn log-uniformly from the stated range. Clean: federations with a false accusation, of 90 and per dataset of 30, with exact 95% intervals that treat the three α cells of a seed as independent. SF, PM: recall at $\rho = 0.2$; gap: median [interquartile range] of the in-sample benign-minus-virtual gap.

Factor range	Clean	5G-NIDD	ToN-IoT	CICIoT2023	SF / PM	Gap
None	0/90; clusters 0/30 [0.000, 0.116]	0/30 [0.000, 0.116]	0/30 [0.000, 0.116]	0/30 [0.000, 0.116]	0.95 / 0.69	−0.697 [−1.071, −0.268]
$[1/1.25, 1.25]$	2/90 [0.003, 0.078]	0/30 [0.000, 0.116]	2/30 [0.008, 0.221]	0/30 [0.000, 0.116]	0.94 / 0.71	−0.427 [−0.878, −0.093]
$[1/1.5, 1.5]$	4/90 [0.012, 0.110]	0/30 [0.000, 0.116]	4/30 [0.038, 0.307]	0/30 [0.000, 0.116]	0.94 / 0.71	−0.411 [−0.876, −0.083]
$[0.5, 2]$	5/90 [0.018, 0.125]	0/30 [0.000, 0.116]	5/30 [0.056, 0.347]	0/30 [0.000, 0.116]	0.94 / 0.71	−0.378 [−0.871, −0.083]

60 clusters. With no event, 36 independent clean federations per dataset are needed for an exact upper bound below $q = 0.1$ (0.097). The 40 federations per dataset at $\alpha = 0.5$ meet this, and the single false accusation among them was one participant in one CICIoT2023 federation.

Table S17 compares the fixed sampler with a sampler matched to the federation. The $\alpha = 0.2$ and 0.3 federations are new partitions of the confirmation seeds, and the 5G-NIDD federation at $\alpha = 0.2$ with seed 10 kept 11 valid participants, hence 359 benign participants. The false accusation at $\alpha = 0.3$ was one 5G-NIDD participant (seed 18), with a cosine score of 0.97 against at most 0.41 for the other eleven. With $\alpha_v = \alpha = 100$ the benign-minus-virtual gap was no longer negative in the leave-one-out score, and the 10 federations with a false accusation were five on 5G-NIDD and five on ToN-IoT. With $\alpha_v = \alpha = 0.2$ the virtual participants scored higher than the benign ones, and recall fell. In the leave-one-virtual-out check no row had more than 0.008 of its p-values below 0.01.

Table S19 Attribution conditional on accusation on the confirmation federations. Attacked and accused: uploads, with distinct attackers in parentheses; hit: accused uploads whose top-attributed feature lies in the edited set. The exact interval treats accusations as independent, and the standardized attribution also has the cluster-bootstrap interval. Spread: accusations whose largest positive standardized deviation holds less than $t = 0.121$ of the positive sum. Edits of the least influential feature were never accused.

Edit	Dataset	Attacked	Accused	Cosine hit	Standardized hit	Std., cluster bootstrap	Spread
Sign-flip, $\rho = 0.1$–0.4	5G-NIDD	360 (150)	341 (150)	341/341 [0.989, 1.000]	241/341 [0.655, 0.755]	[0.665, 0.746]	14/341
	ToN-IoT	360 (150)	337 (145)	337/337 [0.989, 1.000]	217/337 [0.590, 0.695]	[0.540, 0.742]	0/337
	CICIoT2023	360 (150)	343 (147)	343/343 [0.989, 1.000]	155/343 [0.398, 0.506]	[0.352, 0.555]	17/343
Permutation, $\rho = 0.1$–0.4	5G-NIDD	360 (150)	295 (136)	295/295 [0.988, 1.000]	199/295 [0.618, 0.728]	[0.628, 0.727]	14/295
	ToN-IoT	360 (150)	299 (133)	299/299 [0.988, 1.000]	191/299 [0.582, 0.693]	[0.504, 0.759]	0/299
	CICIoT2023	360 (150)	217 (106)	217/217 [0.983, 1.000]	71/217 [0.265, 0.394]	[0.234, 0.417]	13/217
Most influential, $\rho \leq 0.2$	5G-NIDD	180 (60)	131 (54)	131/131 [0.972, 1.000]	122/131 [0.874, 0.968]	[0.875, 0.983]	0/131
	ToN-IoT	180 (60)	171 (60)	171/171 [0.979, 1.000]	147/171 [0.798, 0.908]	[0.772, 0.938]	0/171
	CICIoT2023	180 (60)	127 (57)	127/127 [0.971, 1.000]	102/127 [0.723, 0.868]	[0.688, 0.904]	2/127
Random feature, $\rho \leq 0.2$	ToN-IoT	180 (60)	14 (4)	14/14 [0.768, 1.000]	14/14 [0.768, 1.000]	[1.000, 1.000]	0/14
	5G-NIDD, CICIoT2023	180 (60) each	0	–	–	–	–
Four random, $\rho \leq 0.2$	5G-NIDD	180 (60)	26 (12)	26/26 [0.868, 1.000]	26/26 [0.868, 1.000]	[1.000, 1.000]	0/26
	ToN-IoT	180 (60)	106 (38)	105/106 [0.949, 1.000]	105/106 [0.949, 1.000]	[0.974, 1.000]	0/106
	CICIoT2023	180 (60)	18 (6)	18/18 [0.815, 1.000]	17/18 [0.727, 0.999]	[0.893, 1.000]	0/18
Every feature, $\rho \leq 0.2$	5G-NIDD	180 (60)	180 (60)	–	–	–	138/180
	ToN-IoT	180 (60)	180 (60)	–	–	–	0/180
	CICIoT2023	180 (60)	180 (60)	–	–	–	69/180
Label flipping, $\rho = 0.1$–0.4	5G-NIDD	360 (150)	252 (129)	–	–	–	131/252
	ToN-IoT	360 (150)	54 (25)	–	–	–	0/54
	CICIoT2023	360 (150)	14 (8)	–	–	–	0/14

Table S18 shows that no tested factor range kept the clean federations free of false accusations, and all of them occurred on ToN-IoT. Rescaling raised the benign-minus-virtual gap, and recall was unchanged. Differences between flow exporters beyond per-feature rescaling were not tested.

S10 Attribution conditional on accusation

Table S19 gives the hit rates that an operator receives, conditional on accusation. Accusations of one attacker at nested malicious fractions are not independent, so the cluster-bootstrap interval is the one to read where the two differ. The cosine attribution was right on every accusation of a single-curve edit, and the standardized one on 0.33–1.00 of them per dataset and edit, least often on CICIoT2023. The spread threshold t is the 5th percentile of this share among accused attackers of the 5G-NIDD and ToN-IoT exploration split. Of the 1,080 benign clean confirmation uploads, 0.050 fell below it. The share is bounded below by one over the number of features with a positive deviation, so it depends on the feature count. It marked about half of the label-flip accusations on 5G-NIDD and none on the other two datasets, so it does not identify label flipping.

S11 Colluders that know the audit

The colluders of article Section 7.4 pool every training row they hold. They fit 16 surrogate virtual participants on these rows with the server's sampler ($\alpha_{\mathrm{v}} = 0.5$), their own class presence and their own sample sizes. The surrogate consensus is the median of the surrogate uploads, and the surrogate null uses the same log-t tail. For each feature the edit direction follows the histogram of the colluders' rows over the grid bins, so the edit lands where the global model reads those rows. The targeted direction lowers the $c^\star$ column on the histogram of the colluders' $c^\star$ rows, centred on

Table S20 Damage after excluding the accused for every strategy of the colluders that know the audit (90 confirmation federations per ρ; points; T: targeted, U: untargeted). No audit: largest mean damage over the fixed strengths without the audit; under audit: largest mean damage after exclusion over all strategies, with the strategy; audit adds: their difference. Intervals: paired (dataset, seed)-cluster bootstrap with both maxima re-chosen in each resample. Bold rows: the primary metric of each goal.

				Damage after exclusion, by strategy										
ρ	Goal	Metric	No audit	Cautious	Aggr.	$\lambda = 0.5$	1	2	5	10	20	Under audit (strategy)	Audit adds
0.1	T	accuracy	0.02 [−0.01, 0.05]	0.00	0.03	−0.01	0.01	0.02	0.02	0.02	0.02	0.03 [−0.01, 0.08] (aggr.)	−0.01 [−0.05, 0.02]
	T	$c^\star$ **recall**	0.09 [−0.05, 0.27]	−0.01	0.07	−0.04	0.04	0.03	0.05	0.07	0.07	0.07 [−0.03, 0.21] (aggr.)	0.02 [−0.19, 0.22]
	U	**accuracy**	0.13 [0.06, 0.21]	0.07	0.04	0.05	0.06	0.04	0.02	0.02	0.02	0.07 [0.03, 0.11] (cautious)	0.06 [0.01, 0.11]
	U	$c^\star$ recall	0.25 [0.06, 0.51]	0.11	0.11	0.11	0.13	0.05	0.06	0.06	0.06	0.13 [0.03, 0.27] ($\lambda = 1$)	0.12 [−0.01, 0.29]
0.2	T	accuracy	0.12 [0.03, 0.24]	−0.01	0.02	−0.01	0.02	−0.02	0.00	0.00	0.00	0.02 [−0.01, 0.07] ($\lambda = 1$)	0.10 [0.02, 0.20]
	T	$c^\star$ **recall**	0.82 [0.17, 1.69]	−0.03	0.07	0.02	0.07	0.08	0.15	0.17	0.17	0.17 [0.02, 0.36] ($\lambda = 10$)	0.66 [0.11, 1.36]
	U	**accuracy**	0.50 [0.28, 0.71]	0.17	0.14	0.15	0.19	0.03	0.00	0.00	0.00	0.19 [0.10, 0.29] ($\lambda = 1$)	0.31 [0.16, 0.45]
	U	$c^\star$ recall	0.77 [0.27, 1.35]	0.32	0.27	0.34	0.32	0.12	0.17	0.17	0.17	0.34 [0.11, 0.67] ($\lambda = 0.5$)	0.43 [0.14, 0.74]
0.3	T	accuracy	0.84 [0.36, 1.43]	0.00	0.01	−0.01	0.03	−0.02	0.04	0.04	0.04	0.04 [−0.01, 0.13] ($\lambda = 5$)	0.81 [0.34, 1.34]
	T	$c^\star$ **recall**	5.18 [2.04, 8.99]	0.03	0.04	−0.01	0.09	−0.13	0.08	0.08	0.08	0.09 [−0.02, 0.41] ($\lambda = 1$)	5.09 [2.02, 8.65]
	U	**accuracy**	5.87 [3.83, 7.91]	1.40	1.49	1.27	1.80	0.14	0.04	0.04	0.04	1.80 [1.16, 2.45] ($\lambda = 1$)	4.07 [2.49, 5.74]
	U	$c^\star$ recall	7.20 [3.31, 11.56]	1.33	1.30	1.54	1.13	0.05	0.08	0.08	0.08	1.54 [0.57, 2.76] ($\lambda = 0.5$)	5.66 [2.62, 8.84]
0.4	T	accuracy	2.63 [1.64, 3.60]	0.05	0.09	0.02	0.16	0.04	0.08	0.08	0.08	0.16 [0.08, 0.25] ($\lambda = 1$)	2.47 [1.55, 3.39]
	T	$c^\star$ **recall**	13.79 [7.08, 20.78]	−0.02	0.08	0.02	0.25	−0.08	0.19	0.19	0.19	0.25 [0.02, 0.69] ($\lambda = 1$)	13.53 [7.02, 20.20]
	U	**accuracy**	18.40 [12.82, 23.95]	4.04	4.87	4.34	6.90	0.14	0.08	0.08	0.08	6.90 [4.43, 9.28] ($\lambda = 1$)	11.51 [7.95, 15.41]
	U	$c^\star$ recall	17.74 [9.68, 26.02]	3.63	4.96	4.20	6.18	0.18	0.19	0.19	0.19	6.18 [2.42, 10.32] ($\lambda = 1$)	11.55 [6.54, 17.12]

Table S21 Attacker recall of the real audit (%) for every strategy of the colluders that know the audit, with the mean edit strength λ of the cautious and aggressive colluders. Benign accused: range over strategies of attacked federations in which a benign participant was accused (of 90).

ρ	Goal	Cautious (mean λ)	Aggressive (mean λ)	$\lambda = 0.5$	1	2	5	10	20	Benign accused
0.1	T	3.3 (0.80)	43.3 (3.31)	0.0	4.4	47.8	81.1	91.1	92.2	0
	U	2.2 (0.68)	45.6 (2.34)	0.0	10.0	66.7	96.7	98.9	98.9	0–1
0.2	T	0.0 (0.72)	46.7 (3.48)	0.0	6.1	63.3	93.3	100.0	100.0	0–1
	U	0.0 (0.59)	44.4 (1.78)	0.6	14.4	83.9	100.0	100.0	100.0	0–1
0.3	T	0.0 (0.70)	64.2 (2.57)	0.0	9.4	75.3	100.0	100.0	100.0	0–4
	U	0.0 (0.57)	43.6 (1.63)	0.0	18.9	95.6	100.0	100.0	100.0	0–4
0.4	T	2.2 (0.71)	68.9 (2.97)	0.0	11.1	77.1	100.0	100.0	100.0	0–4
	U	0.9 (0.58)	48.9 (1.82)	0.0	16.4	98.9	100.0	100.0	100.0	0–4

the histogram of all their rows. It raises the column of the class that competes most with $c^\star$ on those rows. The untargeted direction lowers each class held by the colluders on its own rows and raises the next class in turn, and the terms are summed. Each colluder bisects its own λ in $[0, 20]$ against the surrogate audit in 30 steps. A colluder whose honest upload already falls below the threshold sends it unchanged. Damage is measured with the server's intercept refitted on the reference sample, as in every other experiment.

Tables S20 and S21 give every strategy. Damage does not grow monotonically with λ, so without the audit the strongest fixed strength is chosen per cell. Under the audit the best strategy was most often a weak fixed edit ($\lambda = 1$), which the audit accused at recall 0.04–0.19. The server's intercept refit absorbs part of a targeted edit. With the clean model's intercept held fixed, the largest targeted damage after exclusion at $\rho = 0.4$ rose from 0.25 to 2.48 points. Without the audit at $\lambda = 20$ it rose from 13.79 to 16.76 points.

Table S22 Federation against a server-only EBM trained on the server pool: accuracy and macro-F1 in %, differences in points, with (dataset, seed)-cluster bootstrap 95% intervals. Upper block: 30 5G-NIDD federations whose participants hold the flows of one base station; the federation's intercept is fitted on the same pool as the server-only model. Lower block: 15 federations ($\alpha = 0.5$, five per dataset) with pools of 10 and 25 rows per class, tested on the participants' test split.

Server pool	Test rows	Acc., federation	Acc., server	Difference	F1, federation	F1, server	Difference
Other station	participants' station	93.5 [92.0, 94.9]	60.1 [59.8, 60.4]	33.43 [31.94, 34.81]	84.2 [83.2, 85.3]	83.2 [82.5, 83.9]	0.95 [−0.21, 2.17]
	pool's station	96.4 [96.3, 96.6]	99.9 [99.9, 100.0]	−3.50 [−3.67, −3.33]	86.0 [85.2, 86.9]	99.8 [99.7, 99.9]	−13.77 [−14.54, −12.91]
	both	94.7 [93.9, 95.5]	76.1 [75.9, 76.3]	18.64 [17.73, 19.43]	85.5 [84.6, 86.3]	89.8 [89.5, 90.1]	−4.36 [−5.27, −3.45]
Participants' station	participants' station	97.6 [97.4, 97.8]	99.7 [99.7, 99.8]	−2.10 [−2.28, −1.95]	85.8 [84.8, 86.7]	99.0 [98.7, 99.2]	−13.16 [−14.14, −12.33]
	other station	81.5 [80.4, 82.6]	86.1 [82.4, 89.8]	−4.64 [−8.36, −1.20]	80.9 [79.6, 82.3]	94.6 [93.6, 95.7]	−13.77 [−15.53, −11.77]
	both	91.2 [90.7, 91.5]	94.3 [92.8, 95.7]	−3.12 [−4.59, −1.78]	84.5 [83.5, 85.5]	97.7 [97.3, 98.1]	−13.16 [−14.21, −12.11]
10 rows per class (median 90)	test split	90.3 [88.7, 92.1]	81.9 [77.6, 86.4]	8.47 [5.54, 11.24]	77.6 [73.7, 80.9]	72.8 [65.0, 80.3]	4.78 [−0.11, 9.40]
25 rows per class (median 202)	test split	90.5 [89.0, 92.3]	88.2 [85.4, 91.2]	2.37 [0.90, 3.88]	78.2 [74.4, 81.5]	80.4 [74.3, 86.3]	−2.17 [−5.95, 1.23]

Table S23 Accuracy / macro-F1 in clean federations ($\alpha = 0.5$, 10 seeds; centralized models: 5 seeds except on 5G-NIDD) and communication. Upload: $F \times G \times C$ float32 values; MLP rounds and total upload: 20 final runs per dataset.

	MQTTset	5G-NIDD	ToN-IoT	CICIoT2023
Additive, median [ours]	0.869 / 0.510	0.973 / 0.912	0.901 / 0.844	0.906 / 0.651
Additive, mean	0.869 / 0.524	0.973 / 0.925	0.892 / 0.836	0.902 / 0.660
FedAvg MLP (mean)	0.880 / 0.471	0.987 / 0.947	0.932 / 0.862	0.898 / 0.605
FedAvg MLP (coordinate median)	0.876 / 0.468	0.988 / 0.966	0.926 / 0.861	0.905 / 0.609
Federated logistic regression (mean)	0.773 / 0.482	0.928 / 0.640	0.328 / 0.189	0.709 / 0.503
Centralized EBM	0.904 / 0.699	0.998 / 0.995	0.954 / 0.911	0.942 / 0.782
Centralized LightGBM	0.906 / 0.713	0.999 / 0.996	0.966 / 0.930	0.957 / 0.847
Upload, ours (one shot)	15.4 KB	59.0 KB	24.0 KB	49.9 KB
MLP upload per round	43.3 KB	56.9 KB	43.8 KB	55.6 KB
MLP total upload	0.2–3.9 MB	1.4–6.0 MB	2.3–6.6 MB	1.6–4.8 MB

S12 Server pools and server-only training

Table S23 compares the additive federation with federated and centralized alternatives and gives the communication of article Section 7.5.

Table S22 completes the comparison of article Section 7.5. With the pool from the other station, the federation led by 33 accuracy points on the participants' traffic, but its macro-F1 lead was not separated from zero. On the pool's own station the server-only model was ahead on both metrics. With the pool from the participants' station the server-only model led on both metrics on every test set. With pools of 10 and 25 rows per class the federation led in accuracy, but benign ToN-IoT participants were accused in 1 and 2 of 15 clean federations.

Table S24 gives the alternatives to the one-shot federation behind the accuracy comparison of article Section 7.5. The best participant is chosen after the fact, so it is an optimistic reference, and it exceeded the federation on every dataset. The server-only pools are drawn from the participants' own training split and are therefore IID.

S13 Calibration under altered conditions

Table S25 gives every condition of article Section 7.1, of which article Table 4 keeps the rows that the article cites. Under feature shift within $[0.5, 2]$ the five false accusations

Table S24 Alternatives to the one-shot federation: accuracy / macro-F1 ($\alpha = 0.5$, $K = 12$, 10 seeds). Local-only: each participant's own EBM, mean and best participant (the latter chosen after the fact, hence optimistic). Server-only: an EBM trained on a server pool drawn from the participants' training split (S500, S2000, S6000: about 510, 2,000 and 6,000 rows; class-shifted: Dirichlet(0.5) class proportions).

	MQTTset	5G-NIDD	ToN-IoT	CICIoT2023
Federation, median [ours]	0.869 / 0.510	0.973 / 0.912	0.901 / 0.844	0.906 / 0.651
Local-only, mean participant	0.758 / 0.461	0.896 / 0.834	0.767 / 0.690	0.714 / 0.492
Local-only, best participant	0.887 / 0.549	0.993 / 0.947	0.906 / 0.844	0.908 / 0.657
Server-only, S500	0.882 / 0.599	0.990 / 0.969	0.903 / 0.858	0.900 / 0.673
Server-only, S2000	0.892 / 0.610	0.995 / 0.980	0.926 / 0.875	0.924 / 0.719
Server-only, S6000	0.901 / 0.668	0.996 / 0.993	0.941 / 0.900	0.933 / 0.742
Server-only, class-shifted	0.724 / 0.519	0.949 / 0.952	0.858 / 0.814	0.852 / 0.678

Table S25 Calibration and power under every condition of article Section 7.1, with the reference row of each block ($K = 12$ unless stated). Clean: clean federations, or (dataset, seed) clusters, with a false accusation, with exact 95% interval; SF, PM: sign-flip and permutation recall at $\rho = 0.2$.

Condition	Units	Clean	SF	PM
Confirmation (reference)	30 clusters	0/30 [0.000, 0.116]	0.95	0.69
Unseen seeds	30 clusters	0/30 [0.000, 0.116]	–	–
New seeds, $\alpha = 0.5$	60 federations	1/60 [0.000, 0.089]	–	–
All three sets	120 clusters	1/120 [0.000, 0.046]	–	–
$\alpha = 0.5$, all sets: 5G-NIDD and ToN-IoT / CICIoT2023	40 federations each	0/40 [0.000, 0.088] / 1/40 [0.001, 0.132]	–	–
Leave-one-out consensus	30 clusters	0/30 [0.000, 0.116]	0.96	0.77
Feature shift, three features rescaled within $[0.5, 2]$ / $[1/1.5, 1.5]$ / $[1/1.25, 1.25]$	90 federations each	5/90, 4/90, 2/90	0.94	0.71
Quantity skew (participants fit on 25% of their rows)	30 clusters	2/30 [0.008, 0.221]	0.97	0.90
$K = 24$ and 48 ($\alpha = 0.5$, three datasets)	30 clusters	0/30 [0.000, 0.116]	Table S4	
$\alpha = 100$, sampler $\alpha_v = 0.5$ (reference)	30 federations	0/30 [0.000, 0.116]	0.97	0.68
$\alpha = 100$, matched sampler $\alpha_v = 100$	30 federations	10/30 [0.173, 0.528]	1.00	1.00
$\alpha = 0.3$, sampler $\alpha_v = 0.5$	30 federations	1/30 [0.001, 0.172]	0.75	0.48
$\alpha = 0.2$, sampler $\alpha_v = 0.5$	30 federations	0/30 [0.000, 0.116]	0.50	0.30
$\alpha = 0.2$, matched sampler $\alpha_v = 0.2$	30 federations	0/30 [0.000, 0.116]	0.27	0.13
$\alpha = 0.1$, sampler $\alpha_v = 0.5$	10 federations	0/10 [0.000, 0.308]	0.05	0.00
Declared metadata (reference)	45 federations	0/45 [0.000, 0.079]	0.94	0.74
Two colluders declare 5× their count and every class	45	1/45 [0.001, 0.118]	0.98	0.90
No declarations, $\hat{N} = N$	45	1/45 [0.001, 0.118]	0.92	0.84
No declarations, $\hat{N} = 2N$	45	3/45 [0.014, 0.183]	0.98	0.93
Pool of 6,000 training rows (reference)	15 federations	0/15 [0.000, 0.218]	0.87	0.70
6,000 disjoint / 2,000 / about 510 rows	15 each	0/15, 1/15 [0.002, 0.319], 0/15	0.90–1.00	0.60–0.73
Class-shifted (Dirichlet(0.5) class mix)	15	0/15 [0.000, 0.218]	0.70	0.50
25 rows per class (median 202 rows)	15	2/15 [0.017, 0.405]	0.87	0.63
10 rows per class (median 90 rows)	15	1/15 [0.002, 0.319]	0.83	0.63
5G-NIDD, one station: pool from the same station (reference)	30 federations	0/30 [0.000, 0.116]	0.65	0.43
5G-NIDD, one station: pool from the other station	30	0/30 [0.000, 0.116]	0.68	0.50

Note: Intervals treat the units of a row as independent, which holds for clusters and for federations within one α cell but not for federations pooled over α. Each block has its own reference row, and in the first block only the leave-one-out, feature-shift and quantity-skew rows share the reference federations. In the pool block the colluders at $\rho \geq 0.2$ edited the first attacker's feature, so its recalls compare only within the block.

were three of 10 federations at $\alpha = 0.5$ and two of 10 at $\alpha = 1$, all on ToN-IoT. ToN-IoT thus had 5 of 30, with an interval of [0.056, 0.347].

The declaration-free sampler accused in 1 of 45 federations with the true total and in 3 of 45 with twice the total. The single accusations under falsified declarations and under the declaration-free sampler with the true total hit the same benign ToN-IoT participant. With the original virtual participants its p-value of 0.018 was the smallest benign p-value of the confirmation. Both variants drew new virtual participants, so these runs do not separate the effect of the falsified declarations from that of a new draw.

Each of the five standard server pools from the participants' capture left at most 1 of 15 clean federations with a false accusation. The class-shifted pool gave the most

conservative null, with sign-flip recall of 0.70, and 0.30 on 5G-NIDD against 0.80 with the reference pool. The smallest tested pool that kept 0 of 15 false accusations with sign-flip recall of at least 0.6 held about 500 rows, with recall 1.00. The 2,000-row pool had 1 of 15 with recall 0.93. Pools of 10 and 25 rows per class, with medians of 90 and 202 rows, gave 1 and 2 of 15 false accusations, all on ToN-IoT. Their sign-flip recall was 0.83 and 0.87. Fifteen federations cannot separate these counts.

In the test with a pool from the other 5G-NIDD base station, benign flows were 14% of the pool's rows instead of 56% (Section S8). Recall at $\rho = 0.2$ was 0.68 for sign-flip and 0.50 for permutation, against 0.65 and 0.43 with a pool from the participants' own station. The paired differences of 0.03 [−0.08, 0.17] and 0.07 [−0.05, 0.20] include zero. Recall was lower than in the confirmation split under both pool configurations, so the base station of the pool does not explain the difference.

S14 Order of the analyses

The score, the null, the decision rule and the cosine attribution were fixed before the confirmation split was run (article Section 6.3). Every later analysis evaluated this method unchanged, and the following parts were added after earlier results had been read:

- the comparator panel on the 5G-NIDD and ToN-IoT confirmation seeds, after the confirmation;
- the conditions of article Sections 7.1–7.5 beyond the confirmation, such as the unseen and new seeds, feature shift, quantity skew, sampler variants, server pools and knowledgeable attackers;
- the local-only and class-shifted server-only comparisons, after the server-only results had been read;
- two controls that interpret those results, ByMI on IID federations and a pool from the participants' own base station.

S15 Audit algorithm

Algorithm S1 summarizes the audit of article Section 4 for one federation.

S16 Comparators and accuracy baselines

SHAP-on-upload computes the exact mean-|SHAP| importance of each uploaded model on the reference sample and flags robust $z > 3.5$. SHERPA-on-upload applies the procedure of its article [5] to the uploaded models with 30 baseline samples. The two-pass bootstrap null resamples a benign core of participants with robust $z \leq 3.5$ ($B = 200$). The other participant-based nulls are a leave-one-out Student-t null and a normal null fitted to the median and interquartile range of the log scores. This last null is our own quantile-based version of the empirical null, which Efron estimates from a fitted density instead [2]. Sum-L_2 is the position-matched sum over features of $\|\mathbf{S}_j^{(k)} - \bar{\mathbf{S}}_j\|_2$. The per-class score is described in Section S6 and the adaptation of ByMI in Section S7. The FedAvg multilayer perceptron has 128 and 64 hidden units

Algorithm S1 Audit of one federation

Require: uploads $\{\mathbf{S}^{(k)}\}_{k=1}^{K}$; declared class-presence patterns and sample counts; reference pool $\mathcal{R}$; $M = 48$; $\alpha_{\mathrm{v}} = 0.5$; $q = 0.1$

Ensure: accused set $\mathcal{A}$ with cosine and standardized attributions; novelty candidate

exclude all-zero uploads; $\bar{\mathbf{S}} \leftarrow \mathrm{median}_k\, \mathbf{S}^{(k)}$; fit the intercept on the reference sample

$s_k \leftarrow 1 - \cos(\mathbf{S}^{(k)}, \bar{\mathbf{S}})$ for all k (article Eq. 2)

for $v = 1$ **to** M **do**

  draw $\eta_c \sim \mathrm{Beta}(\alpha_{\mathrm{v}}, (K-1)\alpha_{\mathrm{v}})$ for the declared classes of participant $1 + ((v-1) \bmod K)$; sample rows of $\mathcal{R}$ without replacement

  train an EBM; read its curves $\mathbf{S}^{(v)}$; $s_v \leftarrow 1 - \cos(\mathbf{S}^{(v)}, \bar{\mathbf{S}})$

end for

compute p_k via article Eq. (4); $\mathcal{A} \leftarrow \mathrm{BH}(p_1, \ldots, p_K;\, q)$

for $k \in \mathcal{A}$: report $\arg\max_j s_{kj}$ (article Eq. 3) and $\arg\max_j z_{kj}$

novelty candidate $\leftarrow \arg\max_k$ norm of $\mathbf{S}^{(k)}$ on classes held by $< K/2$ participants (review, not accusation)

and quantile-transformed inputs. It trains one local epoch per round for at most 150 rounds, with early stopping on the reference sample. Its input transform and settings were tuned on the seed-0 test accuracy, which favors it. The one-shot federated logistic regression, the centralized EBM and LightGBM models, each participant's own EBM and the server-only EBMs complete the comparison (Tables S23 and S24).

References

[1] Baruch M, Baruch G, Goldberg Y (2019) A little is enough: Circumventing defenses for distributed learning. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019)

[2] Efron B (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association 99(465):96–104. https://doi.org/10.1198/016214504000000089

[3] Fang M, Cao X, Jia J, et al (2020) Local model poisoning attacks to Byzantine-Robust federated learning. In: 29th USENIX Security Symposium (USENIX Security 20). USENIX Association, pp 1605–1622, URL https://www.usenix.org/conference/usenixsecurity20/presentation/fang

[4] Qian C, Wang M, Ren H, et al (2024) ByMI: Byzantine machine identification with false discovery rate control. In: Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 235. PMLR, pp 41357–41382

[5] Sandeepa C, Siniarski B, Wang S, et al (2024) SHERPA: Explainable robust algorithms for privacy-preserved federated learning in future networks to defend against data poisoning attacks. In: 2024 IEEE Symposium on Security and Privacy (SP), pp 4772–4790, https://doi.org/10.1109/SP54263.2024.00271

[6] Shejwalkar V, Houmansadr A (2021) Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning. In: Proceedings of the Network and Distributed Systems Security (NDSS) Symposium 2021, Virtual, https://doi.org/10.14722/ndss.2021.24498, URL https://dx.doi.org/10.14722/ndss.2021.24498