USENIX Security '22 Fall Home

lujianrong@hust.edu.cn ☰

#277 Shielding Federated Learning: Rectifying Direction Is All You Need

Main Edit

Your submissions (All) Search

☑ **Email notification**

Select to receive email on updates to reviews and comments.

▼ **PC conflicts**

None

R2 Reject & Resubmit

Submission (1MB) 13 Oct 2021 4:50:39am PDT · 1af472c7

▼ **Abstract**

Federated learning (FL) enables clients at the edge to collaboratively learn a shared global model without disclosing their private data. Due to its distributed nature, FL is known to be vulnerable to poisoning attacks, where a small number of compromised clients can corrupt the global model by crafting poisoned local models. Although several Byzantine-robust FL schemes have been developed to defend against poisoning attacks, all of them fail to safeguard real-world FL scenarios with heterogeneous and unbalanced data distribution across clients. Moreover, state-of-the-art defenses rely on an auxiliary dataset that has a similar distribution to clients', which obviously violates the privacy principle of FL. It is still challenging to design an effective defense approach that applies to practical FL schemes.

In this work, we provide a deep investigation into poisoning attacks on FL, particularly focusing on the setting where the training data is not independent and identically distributed (non-IID). We first analyze the failures of existing defenses for non-IID data and then propose a general framework to understand the mechanism of poisoning attacks. We then propose HeteroFL, a new Byzantine-robust FL scheme. HeteroFL incorporates four independent defensive modules to make the aggregated model move towards the global optimum associated with no poisoning, by iteratively ruling out updates with particular spatial characteristics. Finally, we design an adaptive attack tailored for HeteroFL to further evaluate its robustness. Our extensive experiments over four non-IID datasets show that HeteroFL

▶ **Authors** (anonymous)

J. Lu, W. Wan, S. Hu, L. Zhang, M. Li, X. Ma, H. Jin [details]

Prior Reviews (PDF, max 15MB) (232kB)

▶ **Topics**

outperforms all state-of-the-art defenses in defeating various poisoning attacks, \eg, HeteroFL achieves similar global model accuracy with the baseline while existing defenses have a 34% to 79% accuracy drop over different non-IID datasets and different models.

	RevRec	RevExp	OveMer	WriQua
Review #277A	3	3	2	2
Review #277B	4	3	4	2
Review #277C	3	2	3	2
Review #277D	1	3	1	4
Review #277E	3	2	2	2

1 Comment: *AuthorFeedback Response* (S. Hu)

You are an **author** of this submission.

Edit submission

Reviews and comments in plain text

Review #277A

Review recommendation

3. Major revision

Reviewer expertise

3. Knowledgeable

Overall merit

2. Top 50% but not top 25% of submitted papers

Writing quality

2. Needs improvement

Paper summary

The paper analyzes the failures of existing defenses for non-IID data and then proposes a new framework to understand the mechanism of poisoning attacks. Specifically, it presents HeteroFL, a new Byzantine-robust FL scheme. HeteroFL incorporates four independent defensive modules to make the aggregated model move towards the global optimum associated with no poisoning by iteratively filtering out suspicious updates. Further, the paper presents an adaptive attack tailored for HeteroFL to further evaluate its robustness. The experiments over 4 non-IID datasets show that HeteroFL outperforms existing defenses in defending against various poisoning attacks.

Strengths

+ The paper considers defending against poisoning attacks on non-IID data, which is a timely topic for federated learning.
+ The paper conducted extensive experiments and ablation studies to evaluate the performance of HetertoFL.
+ The paper considers adaptive attacks to evaluate the robustness of HetertoFL.

Weaknesses

- The 4 defense modules are all based on heuristics and do not provide any guarantees.
- The proposed adaptive attack seems relatively weak. - The writing could be improved and some key details are not clearly presented.

Comments for author

The paper focuses on defending against poisoning attacks in federated learning under the non-IID setting, which is a more practical assumption compared with that used in previous work. Further, the paper proposes two metrics, local diversity and label diversity, to quantify the non-IID degree, which is interesting.

The paper designs 4 independent defensive modules to rule out suspicious updates. However, all these modules are based on heuristics and do not provide any robustness guarantees. Further, some key details are missing. Specifically,

- The Sybil direction removal module considers updates that are extremely close to each other to be potentially malicious, which seems not grounded. Even in the non-IID setting, benign clients may commit similar gradients; thus, this module may greatly affect the normal training in certain scenarios. It is suggested to provide theoretical/empirical support for the rationale of this module.

- The similarity detection module selects an update from each cluster to constitute a new group. However, the paper does not give details about how the update is selected from each cluster. Intuitively, the selection criterion may have a large impact on the effectiveness of HeteroFL.

- The loss evaluation module requires that the server sends the auxiliary data to each client and changes the model architecture, which tends to interfere with the process of federated learning. It is suggested to discuss the feasibility of this module in a practical setting.

- The magnitude rectification module removes updates of too large or small magnitudes and rescales updates to the same magnitude that equals the average of the remaining updates. This module will affect the training efficiency. However, the paper only evaluates the efficiency of HetertoFL under various attacks. It is suggested to evaluate the efficiency of HetertoFL with previous work.

The paper considers an adaptive attack to evaluate the robustness of HeteroFL. However, it only targets the similarity removal module among the 4 defense modules. It is suggested to consider stronger adaptive attacks that consider all 4 four defense modules during crafting malicious gradients. For example, to evade the Sybil direction removal module, the attack should make the malicious gradients far away from each other; to evade the loss evaluation module, it should make sure that the malicious updates result in a small loss on the auxiliary data.

In evaluating the impact of the non-IID degree, the paper uses q to control the non-IID degree. However, the paper doesn't give details about q. In general, different label diversity and load diversity may result in the same q, while the selection of q may favor different defense methods. It is suggested to provide more discussion on the selection of q.

Requested Changes

Pleases see the detailed comments above.

Questions for authors' response

Pleases see the detailed comments above.

Review #277B

Review recommendation

4. Minor revision

Reviewer expertise

3. Knowledgeable

Overall merit

4. Top 10% but not top 5% of submitted papers

Writing quality

2. Needs improvement

Paper summary

This paper looks at poisoning attacks against federated learning, when the data distribution is non-IID. Specifically, it considers both skews in load, and label assignments across the local clients. The proposed approach called HeteroFL, uses a four prong approach to reduce divergence from the true gradient. It first reduces the likelihood of sybil attacks by reducing the impact of similar updates, uses a similarity detection method wherein diverse updates are grouped in order to find those that create dissimilarity, uses a not identical auxiliary dataset to eliminate those updates that do not contribute to loss reduction, and then finally performs rescaling to ensure elimination of biases due to malicious magnitude increases. The paper contains a good evaluation, comparing with other state of the art work, with multiple types of attacks.

Strengths

+ Overall, I found the paper interesting to read, and the approaches thought out.

+ Strong evaluations, showing that the proposed approach works well in being robust to different types of poisoning attacks, and the benefits compared to state of the art approaches.

+ Nice analysis of various components that contribute to the divergence from the true gradient.

Weaknesses

– The analsyis in Section 4.3 and the design of HeteroFL seem disconnected. I wasn't clear how the proposed approaches contribute to the ensuring that the desired $\Delta_d e f^i$ is achieved.

– The paper tries to squeeze in too much of material, with the figures almost illegible. A lot of the material including important things such as the details of datasets have been pushed to the appendix.

– Because of the volume, many of the insights gained do not stand out (e.g., the authors simply gloss over the ablation study -- and the entire paragraph offers no useful take away).

Comments for author

Overall, nice work. The paper however, can be significantly tightened to eliminate redundancies, and text that does not convey much to the reader, to add insights and take aways that do not pop out well.

Review #277C

Review recommendation

3. Major revision

Reviewer expertise

2. Some familiarity

Overall merit

3. Top 25% but not top 10% of submitted papers

Writing quality

2. Needs improvement

Paper summary

The paper introduces a new defence (HeteroFL) against poisoning attacks in Federated Learning. The attack can deal well with non-IID scenarios, including both diversity in quantity per label and labels between clients.

The authors first identify weaknesses of existing defences. Based on these, they define a 4-phase defines: 1) Sybil detection, 2) Exclusion of outlier clusters (likely to be malicious), 3) Exclusion based on loss on a small auxiliary data set and 4) Magnitude adjustment to mitigate harm from remaining malicious updates.

The extensive evaluation indicates that the new defence is more effective than previous defences, even with a specifically crafted attack targeted HeteroFL.

Strengths

+ interesting observation of weaknesses in previous work

+ in-depth evaluation considering a multitude of parameters and settings

Weaknesses

– lack of clarity in many parts of the design (and other parts, but mainly design)

– no real motivation of non-IID scenarios (i.e., why are some realistic and others not)

Comments for author

I really liked the ideas considered in this paper and I am excited by having a privacy-aware defence that deals with non-IID data. I want to see a future version of the paper accepted but at the moment, there is a lot I don't understand about the paper.

Let's start with the motivation of non-IID scenarios (Section 4.1 but also intro). The authors imply that label diversity is more realistic than load diversity but give no evidence. What about if one (rare) label is exclusively owned by one client but others are common for all clients? Is that considered by the non-IID distributions in the paper?

In Sec. 4.2, unless I missed it, the terms \tau and G in Eq. 4 are not defined, so I don't really understand the theorem. A smaller note on 4.3: \tilde{\Delta}^t_i is initially only defined as a deviation for malicious clients, which made me wonder if Eq. 5 only holds for malicious clients. It later becomes clear that it's set to 0 for honest clients but before that, it's hard to follow.

Sec. 5.3 leaves me with a number of questions. Why does the Sybil direction removal call UpdateSelection with two \beta_1s and not \beta_1 and \beta_2? As far as I understand, Sybil detection always removes \beta_2 |C| updates, isn't that problematic if \beta_1 is very different from the fraction of attackers? How do you select the \betas? (I know there is something in the appendix but it did not really add much in terms of explanation, imo) I am wondering if the similarity detection does not cause trouble in case that the situation is IID. I'd expect that then all honest clients end up in one cluster and there may be multiple clusters of different malicious users that are reasonably close together and far from the one benign cluster, resulting in accidentally excluding the benign one. For the loss evaluation, the paper states 'The update whose loss differ- ence with an estimated model is smaller than a threshold λ1'-> do you mean all updates for which that holds or only one? Grammatically, I think it means only one but all such updates makes more sense to me. Furthermore, why is being close to the estimated model an indication of poisoning? I'd expect the opposite.

I struggle a lot with Sec. 6. That might be because I haven't read [12]. I have trouble pinpointing the exact issues, but e.g., why does \mu make the resulting perturbations as similar as possible (last sentence)?

In Sec. 7, I am unsure whether the evaluation includes an IID scenario. My impression is that HeteroFL can have severe problems there (see notes on Sec. 5) and while IID is not the commonly expected scenario for FL, there might still be such scenarios, so handling them could be useful. The key problem is that the parameter q, which determines the degree of non-IID, is not explicitly defined.

Furthermore, while there are many experiments in Sec. 7, explanations are not always given. For instance, can you explain why the most important parts of the HeteroFL (ablation study, A, B, C, D) differ so vastly between data sets? Differences between attacks would make sense to me but it seems like the dataset is more important than the attack...

Editorial quality can be improved, some nits:

- p.1 'of heterogeneous data distribu- tion' -> distributionS
- p.2 comma before which
- p.3 'The central server...obtain' -> obtains
- p.3 size of 4...4 what? Records, bytes?
- p.4 'each type has different amount of data' -> amountS
- p.4 'auxiliary dataset receiving from the server' -> received from the server
- p.5 'more than f gradients' -> I thought it was f or more
- p.5 'Existing study [22] shows' -> An existing study
- p.5 as did in FL -> as done in FL
- p.5 'he aggregated global model in FL suffers from this reduc- tion in accuracy is due to the global model is updated toward an inconsistent point with the optimal point' -> multiple conjugated verbs in one clause
- p.6 'by the same way' -> in the same way
- p.6 ', data poisoning attack in fact only causes' -> article the/a missing before data
- p.6 'that trained under no poisoning' was/has been trained?
- p.8 'Such approach' -> such an
- p.8 'To tackle with this challenge' -> To tackle this challenge
- p.9 what do you mean by 'roundly demonstrate'?

- p.9 'two reason' -> reasonS

Requested Changes

- Improve explanations and descriptions

Questions for authors' response

See above.

Review #277D

14 Jan 2022

Review recommendation

1. Reject

Reviewer expertise

3. Knowledgeable

Overall merit

1. Bottom 50% of submitted papers

Writing quality

4. Well-written

Paper summary

This paper introduces a Byzantine robust SGD for federated learning where half of the clients are adversarial and try to poison the learned model. The paper claims it is still possible to learn in this setting, and introduces a defense that is able to reach high accuracy even with 70% malicious clients.

Strengths

- Well written paper on an important topic that has seen less study than other areas of security+ML work.
- Good discussion that unbalanced datasets are important

Weaknesses

- The paper over-claims in various areas
- No experimental evidence for the flaws of unbalanced distribution on prior papers
- Evaluation has several critical flaws that call the reesults into question

Comments for author

The main concern I have with this paper is that the evaluation has some significant flaws that make the results untrustworthy.

First though let me comment on something somewhat subjective, but I believe it is an important point. This paper claims to have a defense that is "Byzantine Robust". This means something particular: the defense has to be **provably** robust to arbitrary adversaries controlling a significant fraction of clients. Designing an ad-hoc scheme and calling it "Byzantine robust" completely destroys the utility of this word. Under this paper's definition, we might call any distributed system with some attempt at robustness "Byzantine robust".

There's another small writing comment, too. From the start, the paper builds up the defense as if it was some simple method because "Rectifying Direction Is All You Need". However when it actually comes time to develop the defense, "rectifying direction" isn't all that's needed: you first need a sybil detector, then you need to perform direction similarity detection, and finally gradient clipping. Usually the "X is all you need" paper title trend is fairly benign, but in this case, it's actually incorrect. This is something that's easy to fix, so I won't take it into consideration for the ultimate paper decision.

The flaws in the experimental evaluation begin in Section 4.2. The paper makes some very good points about the data distribution not being identical among clients, but then it doesn't actually evaluate this claim made for prior papers. There's some text that hand-waves and says "the median value in each dimension could be significantly different" and this could "result[] in poor performance". I really expect experiments here, though, to show this happens. Otherwise it's just speculation.

My final concerns is with the correctness of the evaluation in Section 7. In Figure 5(a), the authors are claiming that with an "Attacker percentage" of 90% the defense is still robust! This is mathematically impossible, and violates the Byzantine Generals Problem (Lamport et al. 1982). A Byzantine robust system **can not** be robust to more than 50% of its users. And so as a result, the only valid conclusion is that the attacks used here are weak.

But then if we look at Figure 4, it looks like at various points throughout training, the model with the "adaptive attacker" has *better loss* than the baseline non-attacked model. This seems like it should be impossible for any attack.

Figure 7 shows a different questionable situation; \gamma is some form of parameter that controls the strength of the attack (it seems; see below for some questions on this). However when \gamma is increased from 0.03 to 0.09 [edit: corrected from 0.05 to 0.09 after the author response] the attack performs **worse**. This seems like it should never happen: an attack should only get stronger with more power. What's going on here to cause this effect?

(I don't think I fully understand what \gamma is actually doing here. Equation 7 defines \gamma as the arg-max output of some optimization task, but if this is the case, then why are we varying different values of \gamma? And then also why is the search constrained to the rage [0, 0.09]?)

There's one major component of the defense that doesn't make any sense to me. How does the "4 example" held out test set help? From what I can tell, it looks like the adversary is given access to these four examples (the paper says that they're provided to every client, so it seems like they should be) then why can't an adversary just construct a gradient update that makes these four examples have lower loss while still also poisoning the model? It seems like the adaptive attack doesn't take this into account at all, and so in that way it's trivial to detect all poisoned samples because the adversaries never follow the expected protocol.

This would be similar to the the defender saying "please set the 17th-30th bits of the gradient to 0 if you are not evil" and then rejecting the adversarial gradients because they set some of those bits to 1. Yes this would work, but any intelligent adversary would just follow the protocol. I would expect an evaluation that does at least this.

In Table 1 and Table 2, the "Adaptive attack" success rate on HeteroFL is **weaker** than the "SF attack" for CIFAR-10. This is counterintuitive: it means that the attack *specifically designed to break this defense* is not doing as good of a job at attacking the defense when compared to the trivial attack of just flipping the sign. Why should we trust the results of this adaptive attack?

Table 1 is not possible to interpret without stating the fraction of clients who are adversarial, but this detail appears to be missing. Trying to match the data with Figure 5 makes it look like it's at a ration <10%, because in this figure the CIFAR-10 model has <70% accuracy at a 10% malicious client rate but the value in Table 1 is 71%. What value is used here? (It's possible I missed this number, but it's not listed anywhere around the table or in in Section 7.)

The ablation study is interesting because it appears to show that none of the pieces of the defense are that important by themself. In particular, even if we remove any of the pieces, the model accuracy will only drop by <5% more. Why not just design a simpler defense then that doesn't use the components that aren't important, instead of complicating the defense with extra pieces?

There are then a number of much more minor concerns:

- I don't understand why rectification changes the angle of the vector in figure 3(d), shouldn't it just alter the length?
- Why is it not possible to fool the sybil detector by making the gradients have distance just outside whatever the detection threshold is?
- The models that are trained are very low accuracy to start out with. A ResNet on CIFAR-10 should reach 90%+ accuracy, a 70% accuracy is too low to be interesting---and likewise for the other datasets.
- The authors should release source code of this defense along with the paper; the defense is sufficiently complex I don't believe that by reading the paper someone could faithfully reproduce the claims.

The authors claim that this scheme is indeed robust in the presence of 90% of malicious users is deeply worrying. I strongly recommend the authors read literature on byzantine robustness in order to understand how incredible (here taking the definition "impossible to believe") this claim is. If the authors truly believe that they are able to achieve such a strong result, they should make this claim the main focus of the paper.

The rest of the concerns are all secondary until this problem is resolved. The fact that the adaptive attack performs, in some cases, less strongly than the sign flipping attack is an indicator for what might be going wrong. As is the fact that increasing lambda makes the attack perform less well. But these are just diagnostics that hint at the core issue: the adaptive attack is likely far weaker than a true strong attack.

Review #277E

Review recommendation

3. Major revision

Reviewer expertise

2. Some familiarity

Overall merit

2. Top 50% but not top 25% of submitted papers

Writing quality

2. Needs improvement

Paper summary

The authors provide a detailed understanding on why poisoning attacks occur in the non-IID setting. Using this knowledge, they provide a multi-stage defense mechanism to help remove data poisoning effects in federated learning. This mechanism relies on four components. The first component performs sybil detection to remove similar updates, which could come from the adversary. The second component creates clusters of gradient updates that belong to similar data distributions and removes the outlier updates. The third component performs loss evaluation on an auxiliary dataset. The fourth component performs magnitude rectification by forcing all gradient updates to be within the same order. The proposed approach is evaluated on several datasets, where it can detect poisoned updates, even against adaptive attacks.

Strengths

- The authors consider an adaptive analysis, which is often skipped in prior work.
- The evaluation considered in the paper is exhaustive.

Weaknesses

- The design aspects of the proposed approach are not well-justified theoretically
- The proposed approach might reduce accuracy in IID settings
- The paper could be improved in terms of writing, verbosity, formalism, and notation.

Comments for author

This paper presents an interesting contribution to protect against poisonous updates in the non-IID setting of federated learning. In that setting, it is more challenging to discern anomalous updates as poisonous updates are technically out of distribution. If the setting is non-IID, the adversary can leverage the situation to disrupt the training process. These are a few issues that the paper does not address clearly.

First, the proposed defense mechanism seems like a combination of several prior works; loss evaluation borrows inspiration from Zeno etc., magnitude rectification borrows insight from trimming. This reduces the novelty of the proposed approach. Could the authors comment on how they define their actual contribution relative to related works?

Second, there are some design decisions that are not well-justified. For example, the sybil detection removal relies on the premise that similar updates in the non-IID setting are from malicious clients. But this depends on the degree of non-IIDness, which is not known apriori. Could the authors comment on this aspect?

Further, this approach could easily drop the accuracy of the model in the IID setting. It is not clear how the defender would identify the IID-setting while little less than half of the gradient updates can be compromised. Another example is similarity detection, where the security intuition is not clear. For example, it is not clear why “Each centroid can be regarded as a gradient that is trained over a dataset that covers samples as diverse as possible. Therefore all centroids can be viewed as new updates trained in an IID-like setting.” It is not clear how this would thwart an attacker countrolling 49% of the gradient updates. The paper contains no theoretical argument to this effect.

Moreover, the description of the loss evaluation is not clear. There are two questions: Is D_{aux} the same as the public dataset that’s used for loss evaluation? Also, why are only 4 samples sufficient? The paper does not appear to provide an insight provided for this number 4.

Third, section 4.3 – which is key to the paper has some issues in presentation. First, $d_b^{t,l}$ is used to denote the optimal gradient for the b^{th} batch (which I am assuming comprises multiple samples, since the batch size is S). Yet, the term is calculated for a particular sample (x,l). Could the authors clarify what this means? Does this mean that the batch size S is always equal to 1? A similar issue applies to $g_{i,b,e}^{t,l}$ as well.

Also, the deviation for a particular client is calculated as the difference between the optimal and observed gradient for all batches. For the accumulated deviation of client i, why is there a summation over all labels? The intuition behind this is not clear.

Fourth, the paper could use improvements in its writing, as follows:

- Describing the datasets in the appendix is not ideal. As the paper is verbose in places, some sections could be summarized to make space. Further, some of the results can move to the appendix.
- Several statements made in the paper can come across as imprecise. For example, what do the authors mean when they say that non-IID setting “enlarges” the legitimate direction space? Other examples are “significantly damage the global model” and “slightly change the direction ..” These notions are not formalized.
- It would make for an improved reading experience if the text describing the algorithm and the algorithm(s) are co-located. This is the case throughout section 5.3.

Finally, there are items that require clarifications:

- What is the AGR-tailored attack framework?
- What are assumptions 1 and 2 in theorem 1?
- What does data t refer to in the explanation of Zeno?
- What does it mean for the updates to be similar to each other in S 4.2?
- What does X_i (i \in [n]) used to represent in S 4.3? The data distribution for each of the n clients?

AuthorFeedback Response

Author [Shengshan Hu] 13 Jan 2022 691 words

Q: Performance of sybil direction removal module in IID setting. (Rev#277A-C1, Rev#277C-C4, Rev#277E-C2)

Re: We must emphasize that when we use cosine similarity (not Euclidean distance) as the evaluation metric, the updates from benign clients are unlikely to be similar even in the IID setting. This is because when the model converges, (benign) gradients will gradually become zero, thus making their cosine similarities close to zero, which indicates that those gradients are orthogonal and dissimilar from each other. This phenomenon has been observed in our experiments and will be added to our revised paper. Therefore, this module will not, at least with a low probability, discard benign updates in IID or non-IID setting.

Q: Details of similarity detection module. (Rev#227A-C2, Rev#277C-C4, Rev#277E-[W2, C3])

Re: We construct the group simply by randomly selecting an update from each cluster. We currently do not consider other complicated selection criteria. This grouping step can solve the clustering trouble in the IID setting (e.g., accidentally excluding the benign one), as it redistributes updates from the same cluster to different groups. We note that this module cannot identify malicious updates when there are many attackers (e.g., 49%), a problem which is addressed by other modules.

Q: Consider adaptive attacks against four modules. (Rev#277A-C5, Rev#277D-C9)

Re: Our experiments have actually considered adaptive attacks against all four defense modules rather than just the similarity removal module. AGR-tailored attack (Eq.~(7)) is a general framework targeting the final defense results regardless of how many intermediate defense modules were used. In Eq.~(8), we instantiate the attack by setting C to be final updates after performing four defensive steps (ie, $C = C_4$).

For the loss evaluation module, Tables 1&2 show that under adaptive attack, the model accuracy has a slight reduction. This shows that there exist some malicious updates that have evaded the detection of this module, which may be the attacks proposed by Rev#277D-C9.

Q: Lack of theoretical guarantee. (Rev#277A-W1, Rev#277E-[W1, C3])

Re: The theoretical security analysis is moved to Appendix G.

Q: Details about loss evaluation module. (Rev#277E-C4)

Re: D_{aux} denotes samples used for loss evaluation. We experimentally found that 4 samples are sufficient since using trained data to test the corresponding model makes it easy to find abnormality.

Q: Novelty clarification (Rev#277E-C1)

Re: HeteroFL has three novel ideas. Firstly, it propose a clustering-then-grouping strategy to improve the IID degree of updates. Secondly, it exploits auxiliary dataset without breaking privacy guarantee. Thirdly, it proposes a new direction rectification method.

Q: Misunderstandings on experimental results. (Rev#277D-[C6, C7, C11])

Re: Fig. 4 shows the global model testing loss, a larger loss indicates a more powerful attack.

In Fig. 7 when γ is increased from 0.03 to 0.05, the accuracy drops in four datasets, showing that the attack performs better.

Table 2 shows adaptive attack incurs a larger accuracy reduction (i.e., 0.0098) than SF, while SF incurs a larger reduction (i.e., 0.0003) in Table 1. We conclude that these two attacks perform comparably under some particular cases.

Q: Misunderstanding on γ. (Rev#277D-C8)

Re: Fig. 7 only evaluates the impact of the initial γ, rather than changing the arg-max outputs of Eq. (8). Fig. 7 only demonstrates the results where initial γ is set to the range [0, 0.09], because the model accuracy is reduced close to 0 when the initial γ is larger than 0.09.

Q: Concerns on the correctness of evaluations with 90% attackers. (Rev#277D-C5)

Re: In traditional Byzantine problems, it is true that the system cannot be secure when more than 50% users are malicious. In the FL scenario, however, we can construct some trusted priori knowledge (i.e., the auxiliary dataset D_{aux}) to aid the server to identify all the malicious updates that behave abnormally over these priori knowledge. Therefore Zeno, FLtrust, and our HeteroFL are still robust when there are 90% attackers (shown in Fig. 5(a)).

Q: Why not design a simpler defense? (Rev#277D-C13)

Re: Although the accuracy is reduced by 5% after removing one module, it is very important and difficult to improve the accuracy by 5% in many scenarios. Besides, each module is computation-efficient, it is better to construct a scheme with a stronger security guarantee.