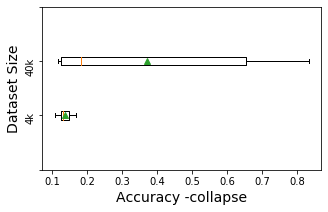
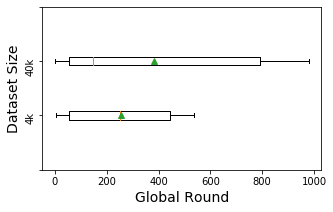
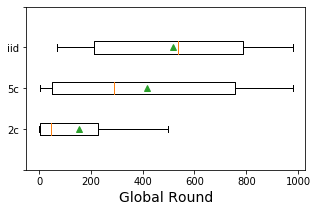
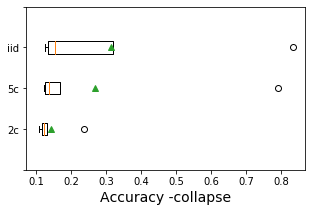
Poisoning Model

We identify the performance to be fully collapsed when the accuracy decreased to its lowest point after the minimal loss (a threshold of 0.025 was applied to eliminate the influence from fluctuation). If accuracy does dot collapse within 1000 global rounds, collapse index is set as 1000.

In figure xxx and figure xxx below, we show the speed and the severity of collapsing (using ‘x%-1’ results only), grouped by dataset size or non-IID levels. Though we can observe increasing accuracy at early training rounds in most cases, these rounds are counted in to collapsing rounds since they are also affected by poisoning weights. Thus, collapsing speed is represented via the round index of collapsing point. Collapsing accuracy is the accuracy at that round, it represents the worst performance of the model in 1000 rounds.

We assume training with large or IID dataset will collapse at slower speed than with small or non-IID dataset, and its accuracy at collapse point will also be higher. Statistical tests have been applied between each IID groups or dataset size (exclude ‘3%’ results when comparing between dataset sizes). Unfortunately, there is not significance shown between IID levels or dataset sizes.

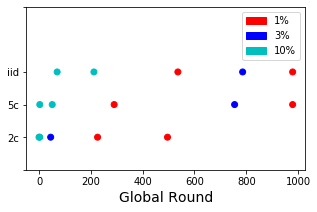
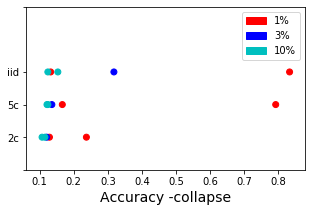
 

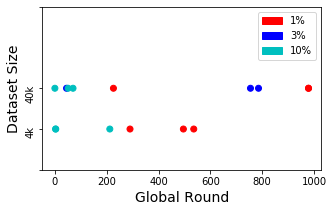
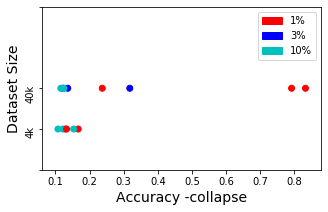


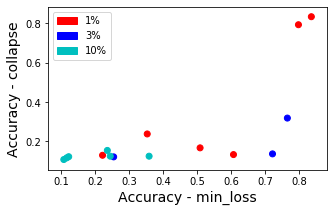
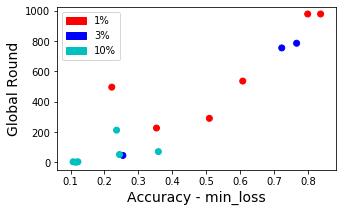
(in dataset size comparison, ‘3%’ is neglected)

We also calculate Pearson correlation coefficient between model’s best performance (accuracy at minimal loss) and collapse speed (index of collapse round). The result shows a positive correlation (0.92, p-value = 1.33e-06). Positive correlation also appears between model’s best performance and the accuracy at collapse round (0.70, p-value = 0.0035). It means seriously affected trainings or less robust models will collapse at faster speed and to worse results.

***Model Poisoning***







***Model Poisoning***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ‘4k’-‘40k’ | ‘IID’-‘5c’ | ‘5c’-‘2c’ | ‘iid’-‘2c’ | ANOVA |
| Acc\_collapse | 0.14# | 0.059# | 0.42# | 0.059# | 0.54 |
| Round\_collapse | 0.67# | 0.25 | 0.13 | 0.085 | 0.29 |

