A LITTLE IS ENOUGH: CIRCUMVENTING DEFENSES FOR DISTRIBUTED LEARNING

Moran Baruch, Gilad Baruch, Yoav Goldberg

Bar Ilan University, Israel

BIU NLP

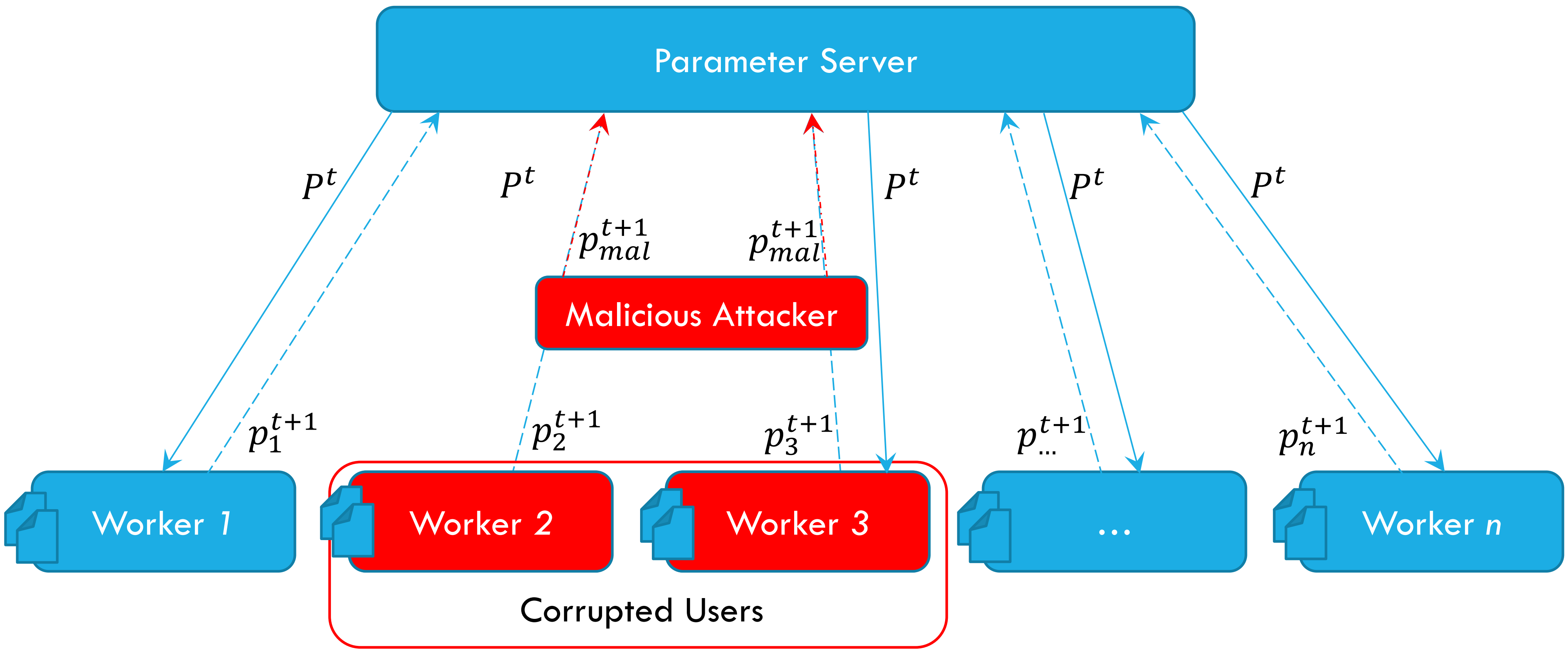
Bar-Ilan University

Tradition of Excellence

BIU

Center for Research in Applied Cryptography and Cyber Security

ATTACKING DISTRIBUTED LEARNING



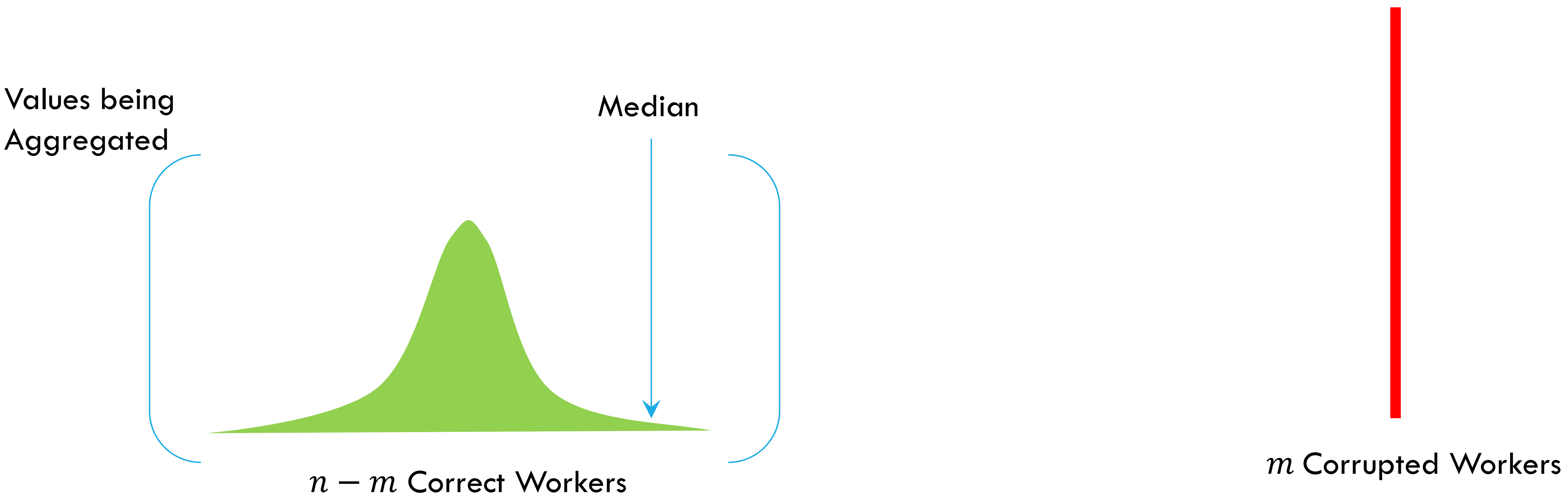
STATISTICS BASED DEFENSES

- Assumptions:
  - The different chunks are assumed to be i.i.d
  - Any attack will require large changes
  - The variance between correct workers is low

Using statistics to discard “outliers”

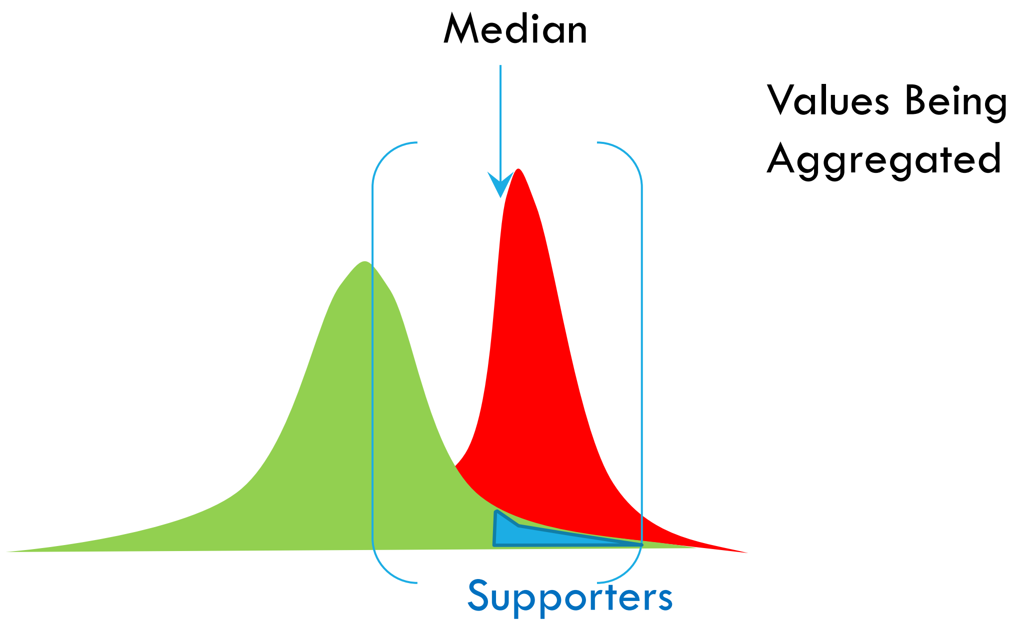
DEFENSE EXAMPLE – TRIMMED MEAN

- Working on each dimension separately
- Finds the median and aggregate only values close to it



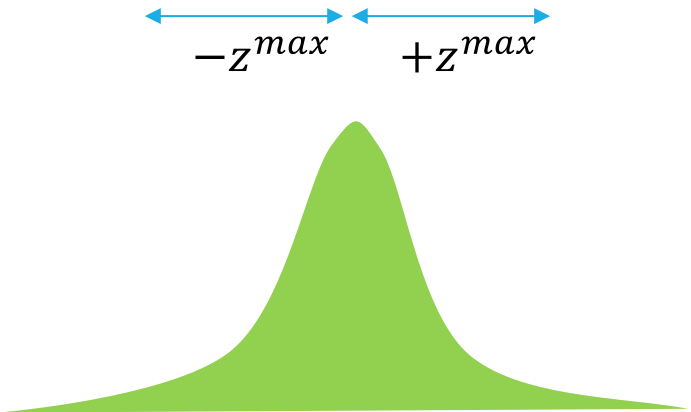
ATTACK MOTIVATION

- There are correct workers with extreme values
  - Use those workers as “supporters” for the change we want to apply
- Apply **small changes, on each dimension,** that will prevent robust statistics



OUR ATTACK

- Use $\phi(z)$ to find maximal allowed change z^{max} that will not be detected
  - Units of standard deviation
- We show that the standard deviation between correct workers are big enough to allow:
  - Prevention of convergence
  - **Backdooring the model**



EXPERIMENTAL RESULTS

Convergence Prevention:

- Our attack was able to reduce the accuracy using the same attack configuration
  - 30-50% degradation for models trained on CIFAR10 and CIFAR100
  - 8-18% degradation for the model trained on MNIST

Backdooring

- The backdoor was introduced correctly in all models
  - Less than 7% degradation in accuracy on benign inputs for MNIST and CIFAR10
  - Up to 20% degradation in accuracy on benign inputs for CIFAR100

STOP BY OUR POSTER!