

Defenses?

Selective gradient sharing

Dataset: Text reviews

Main Task: Sentiment classifier

Doesn't really work...

Participant-level differential privacy

Hide participant's contributions

Only 2 “hand-crafted” mechanisms in the literature

Fail to converge for “few” participants

Property / % parameters shared	10%	50%	100%
Top region	0.84	0.86	0.93
Gender	0.90	0.91	0.93
Veracity	0.94	0.99	0.99

Defenses?

Selective gradient sharing

Dataset: Text reviews

Main Task: Sentiment classifier

Doesn't really work...

Property / % parameters shared	10%	50%	100%
Top region	0.84	0.86	0.93
Gender	0.90	0.91	0.93
Veracity	0.94	0.99	0.99

Participant-level differential privacy

Hide participant's contributions

Only 2 “hand-crafted” mechanisms in the literature

Fail to converge for “few” participants

Thank you!