Defenses?

Selective gradient sharing

Dataset: Text reviews

Main Task: Sentiment classifier

Doesn't really work...

Participant-level differential privacy

Hide participant's contributions

Only two 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

42

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 two mechanisms in the literature
- Fail to converge for "few" participants

Membership Inference

Yelp-health		FourSquare	
Batch Size	**Precision**	**Batch Size**	**Precision**
32	0.92	100	0.99
64	0.84	200	0.98
128	0.75	500	0.91
256	0.66	1,000	0.76
512	0.62	2,000	0.62

Two-Party Membership Inference
(Recall is 1.0)