Private Multi-Party Machine Learning in an Untrusted Setting

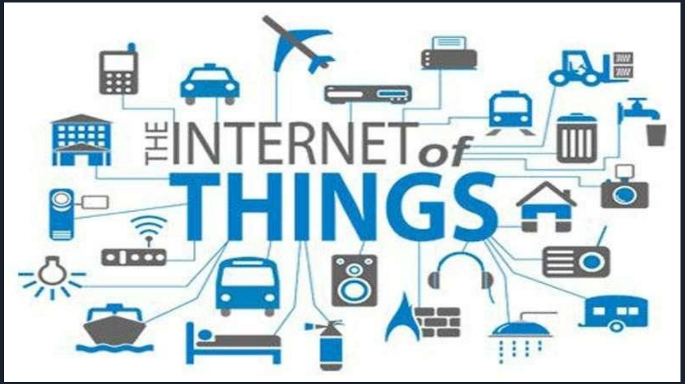
Clement Fung, Jamie Koerner, Stewart Grant, Ivan Beschastnikh

University of British Columbia

Networks Systems Security lab
http://nss.cs.ubc.ca

Data and Analysis are Decentralized

- Internet of Things (large scale sensor networks)
- Live mobile analytics (maps/routing/traffic)



Google Maps

Centralizing Data is a Concern: Privacy

- Data can be sensitive in nature
  - Photos, location info, voice recordings
- Typically, a centralized service performs model training
  - Do we have to trust Google with our data?

Centralizing Data is a Concern: Privacy

But, Privacy is Difficult

- 2006 Netflix user dataset de-anonymized using IMDB [1]
- 2006 AOL search database de-anonymized [2]
- <u>Anonymizing is insufficient</u>: auxiliary data breaks anonymity!

...

[1] Narayanan et al. "*Robust De-anonymization of Large Sparse Datasets*", S&P '08
[2] NYTimes "*A Face Is Exposed for AOL Searcher No. 4417749*" NYTimes '06

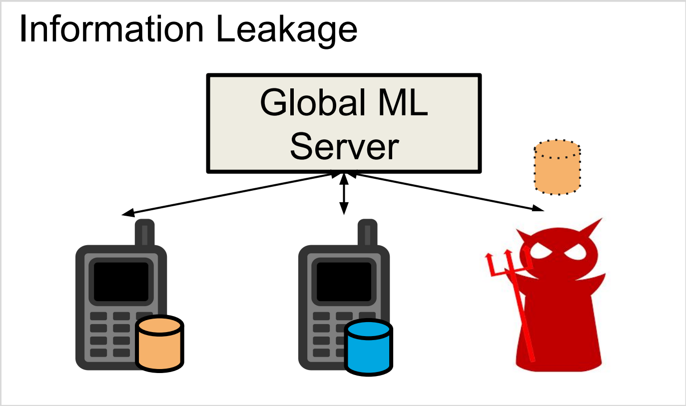
Our Approach: Brokered Learning

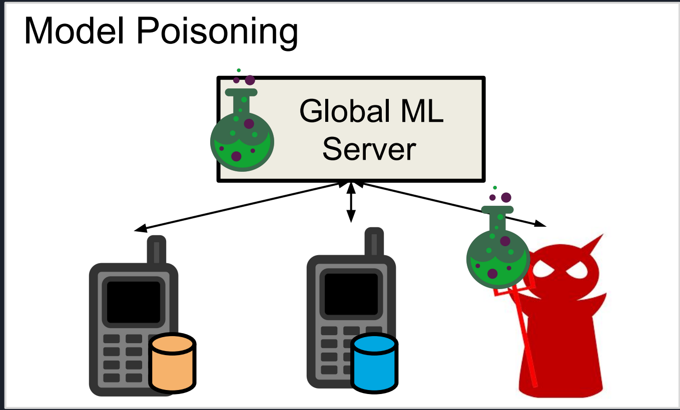
- **Problem:** Current collaborative ML solutions rely on an unsophisticated threat model: Trust the central service
- **Our solution:** New brokered learning model for privacy-preserving anonymous ML
- New defences against known ML attacks for this setting

TorMentor: A system for private, anonymous ML

Attacks in Current ML Architectures

- **Despite trusted infrastructure**, dishonest clients are a threat!
  - Steal your training data
  - Compromise your model

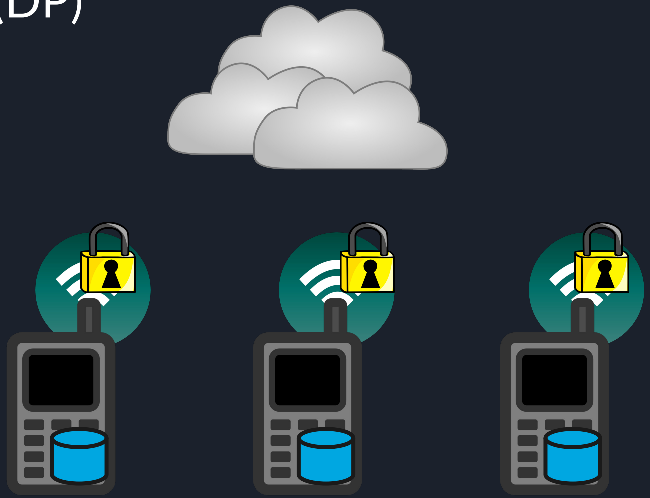




[1] Hitaj et al. "*Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning*" CCS '17
[2] Huang et al. "*Adversarial Machine Learning*". AISec '11

Client Centric ML Defenses

- Modern defenses are not robust to these attacks!
  - In distributed ML, these must be **client centric**
- Current state of the art for client centric defenses:
  - Differential Privacy (DP)
  - Anonymity

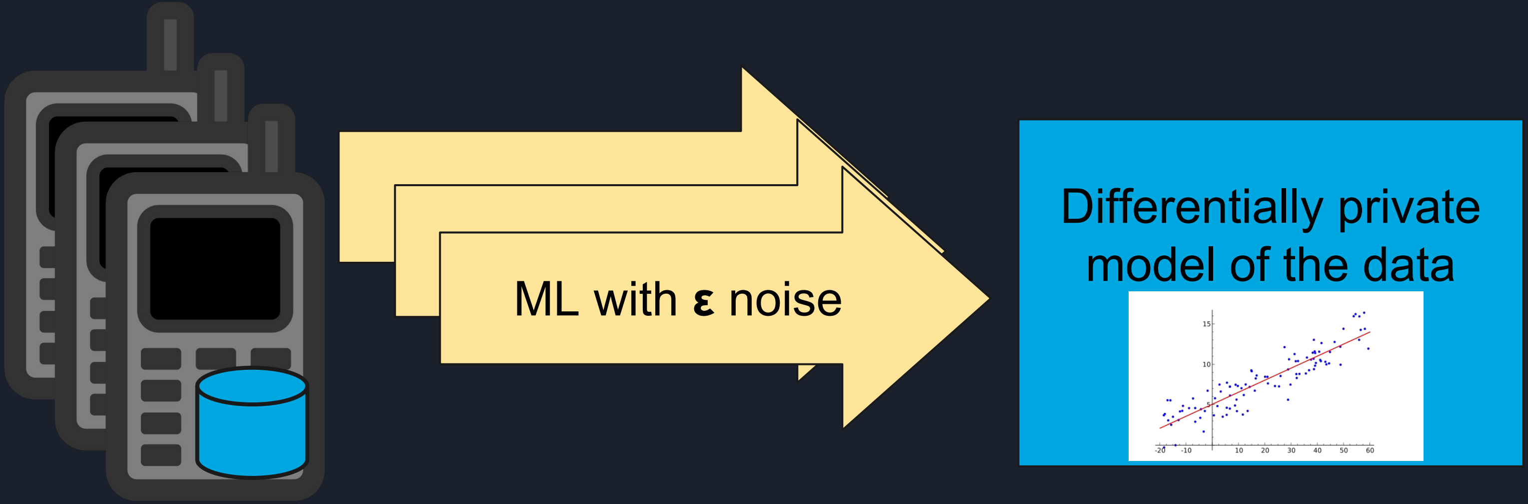


Differential Privacy (DP) in ML

- In ML, DP used to protect training data privacy
  - Applied in SVM, random forest, deep learning, etc.
  - Differentially private SGD: client-side protection
    - Apply parameterized noise (**ε**) to SGD updates
    - More noise leads to worse models

[1] Cynthia Dwork. "*Differential Privacy*" ICALP '06
[2] Song et al. "*Stochastic gradient descent with differentially private updates*" GlobalSIP '13
[3] Oravec et al. "*Efficiency of Recognition Methods for Single Sample per Person Based Face Recognition*" Reviews, Refinements and New Ideas in Face Recognition. 2011.

Differential Privacy (DP) in ML



[1] Cynthia Dwork. "*Differential Privacy*" ICALP '06
[2] Song et al. "*Stochastic gradient descent with differentially private updates*" GlobalSIP '13

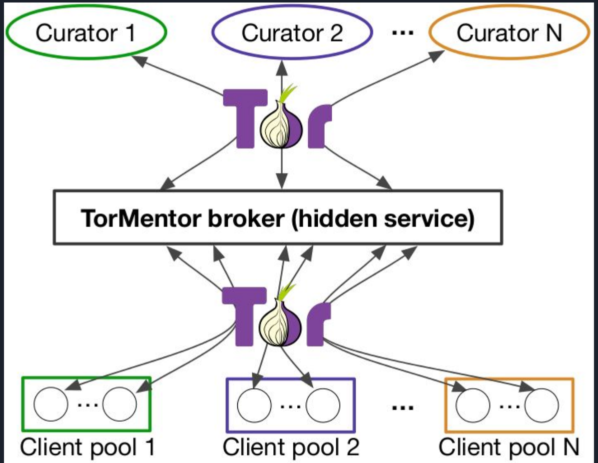
Tor: Anonymity in P2P Systems

- As client: avoid being targeted in information leakage attack
- Use onion routing protocol (Tor)
  - Communicate through chain of random nodes in system
  - Can hide identity of clients in distributed ML!

[1] Dingledine et al, "*Tor: The Second-Generation Onion Router*" Usenix Sec '04]

TorMentor: Putting It All Together

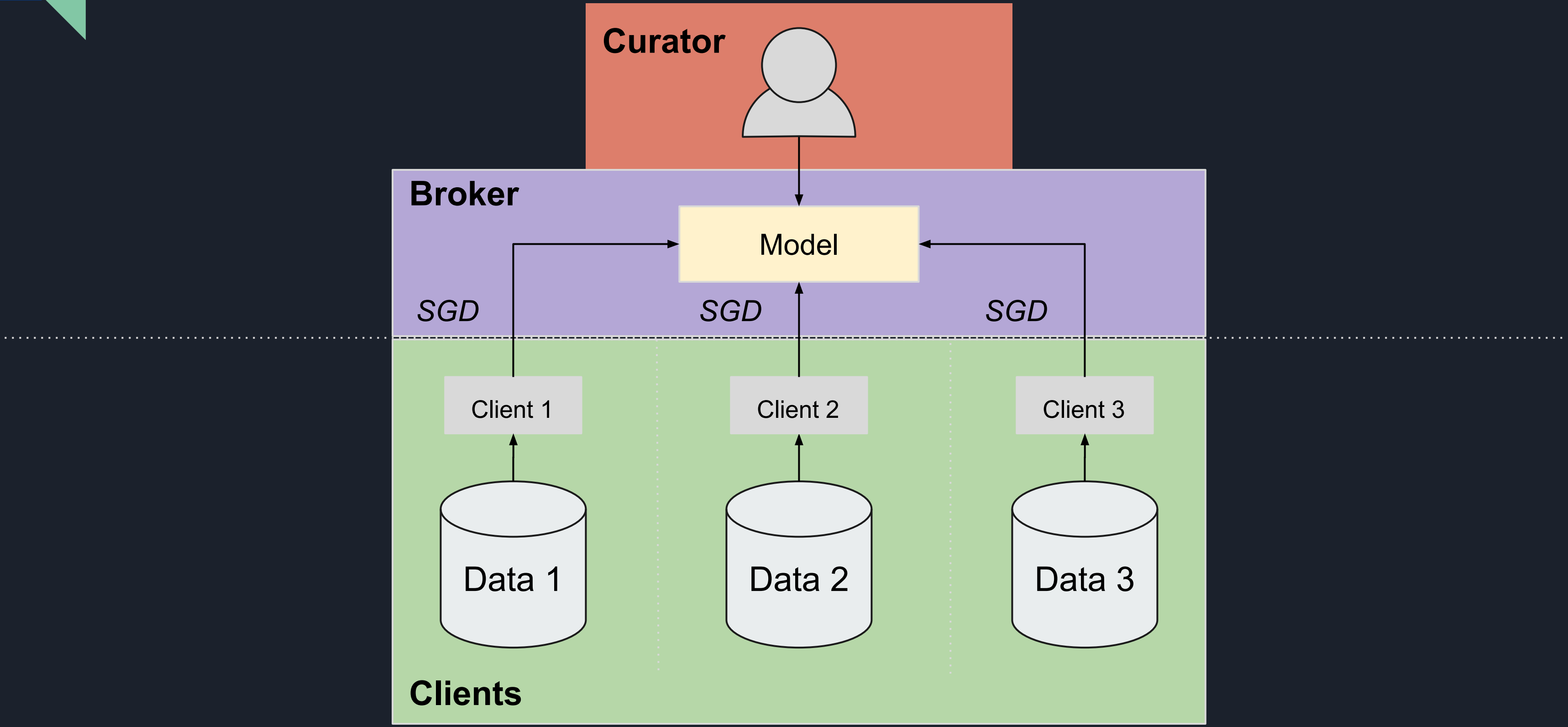
- Data at scale: decentralized ML systems
  - Federated learning, Gaia
  - **But these are not privacy focused**
- Privacy: differentially private ML
  - Differentially private SGD
  - **But these are theoretical**
- Anonymity in P2P systems
  - Onion Routing with Tor
  - **But how can it be used to coordinate ML?**



[1] McMahan et al. "*Communication-Efficient Learning of Deep Networks from Decentralized Data*" AISTATS '17
[2] Hsieh et al. "*Gaia: Geo-Distributed Machine Learning Approaching LAN Speeds*" NSDI '17

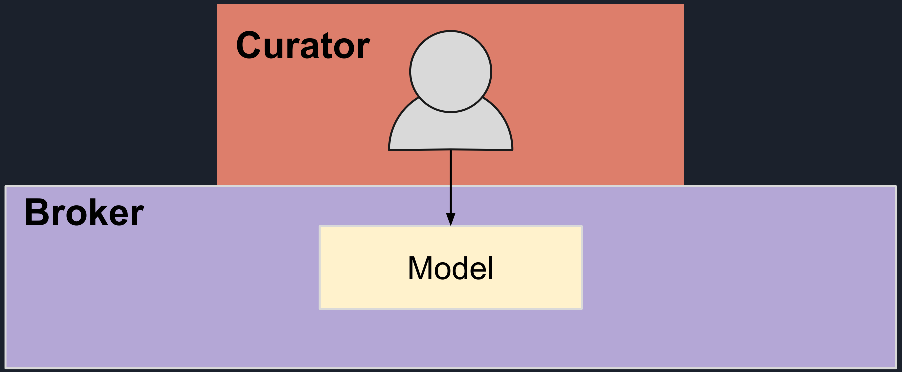
Brokered Learning

Brokered Learning



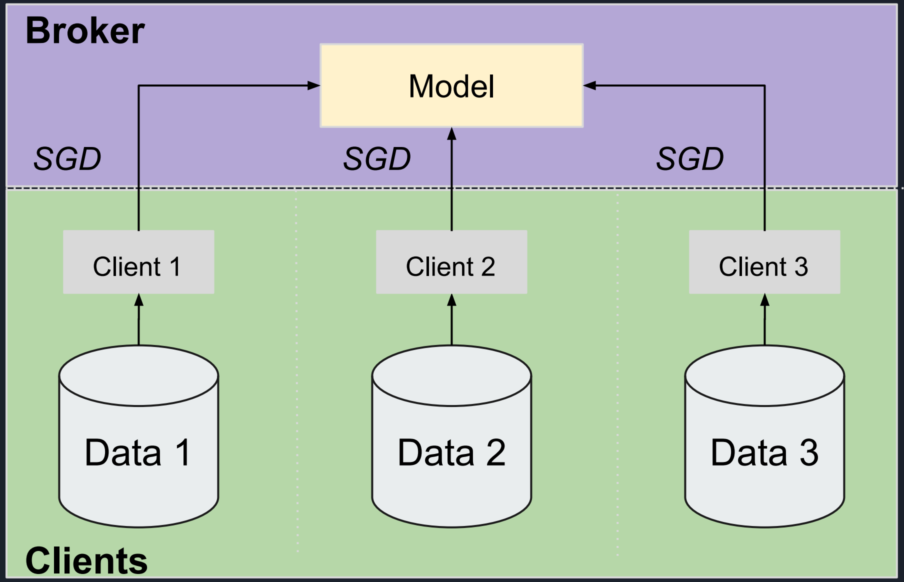
Brokered Learning: Curators

- Decouple model definers and infrastructure providers



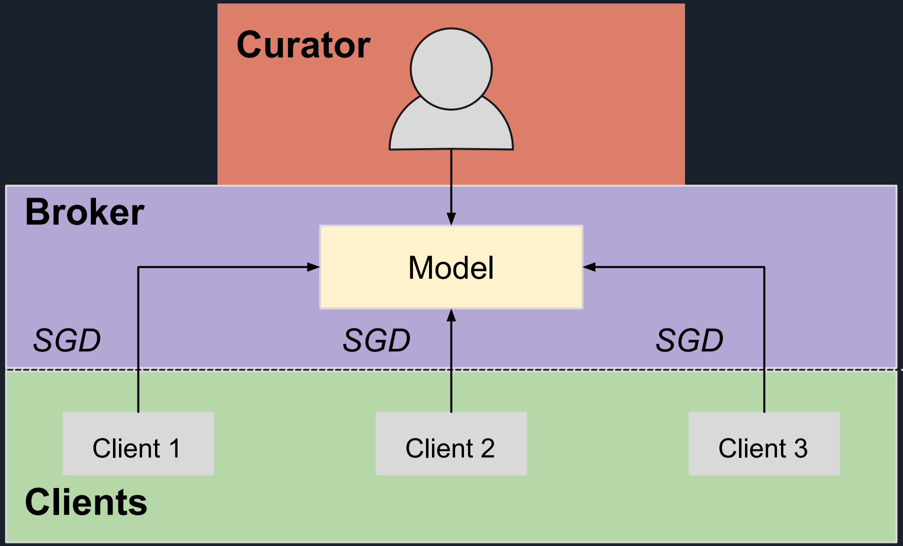
Brokered Learning: Clients

- Allow users control over their own privacy levels ε
- Opt-in, opt-out of model training anytime

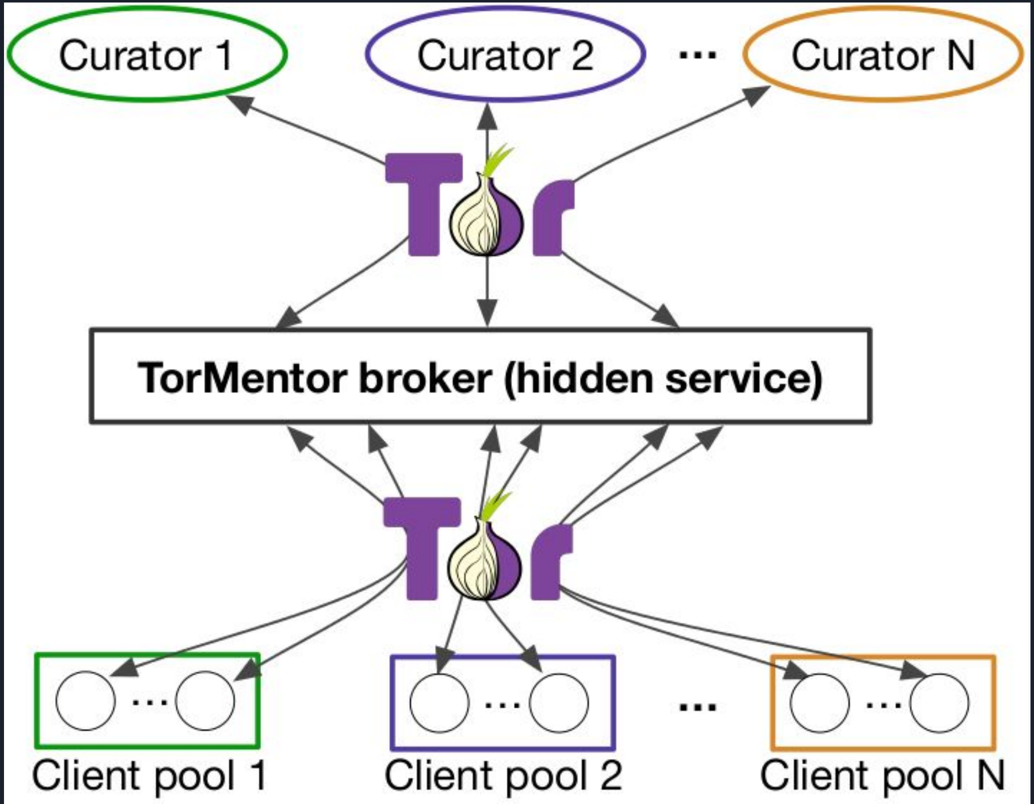


Brokered Learning: Broker

- Provide trusted guarantees to facilitate anonymous ML



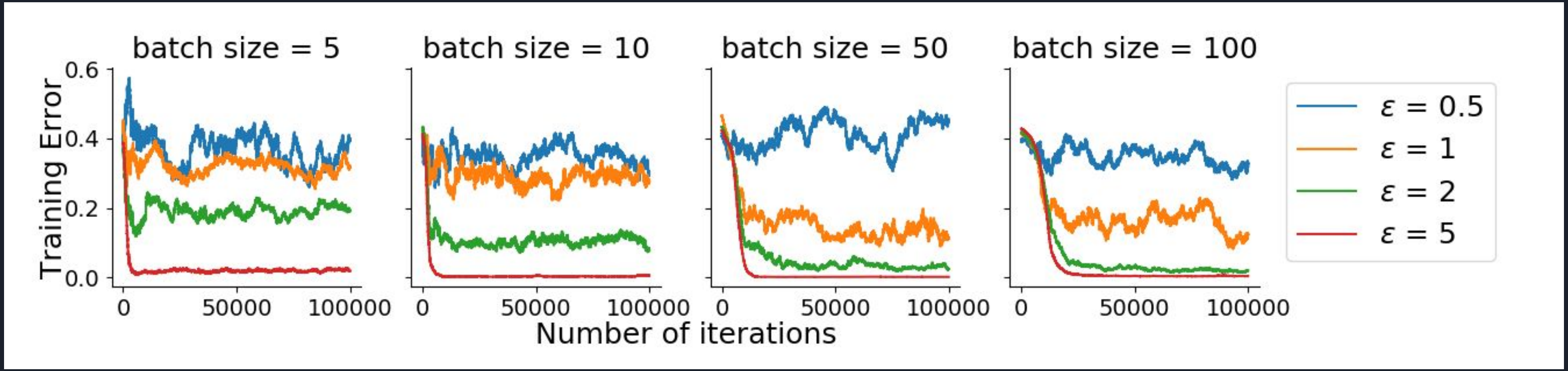
System Overview



Results

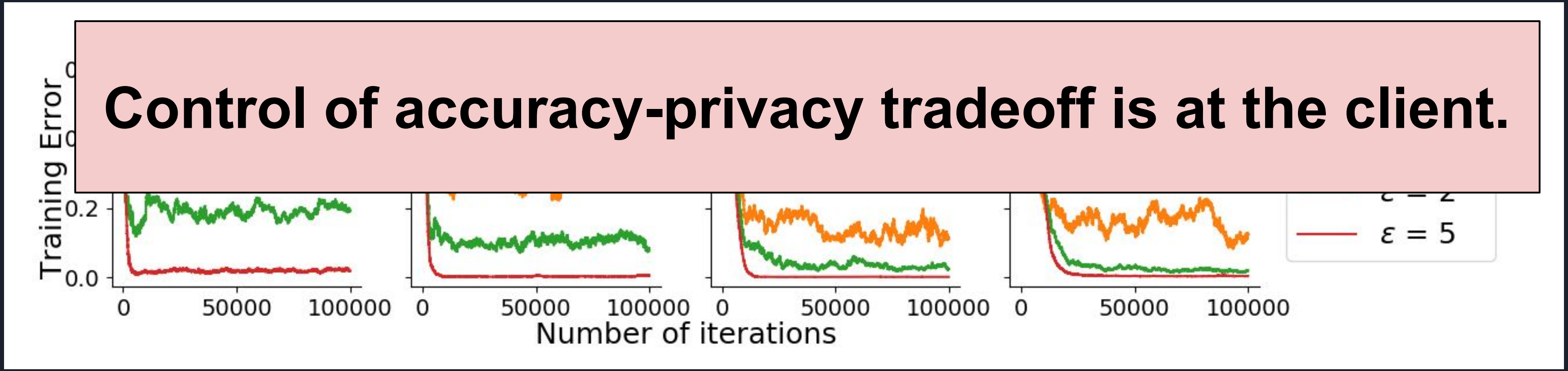
Effect of Client-Side Privacy Parameters

- Varying batch sizes and privacy parameters in TorMentor



Effect of Client-Side Privacy Parameters

- Varying batch sizes and privacy parameters in TorMentor

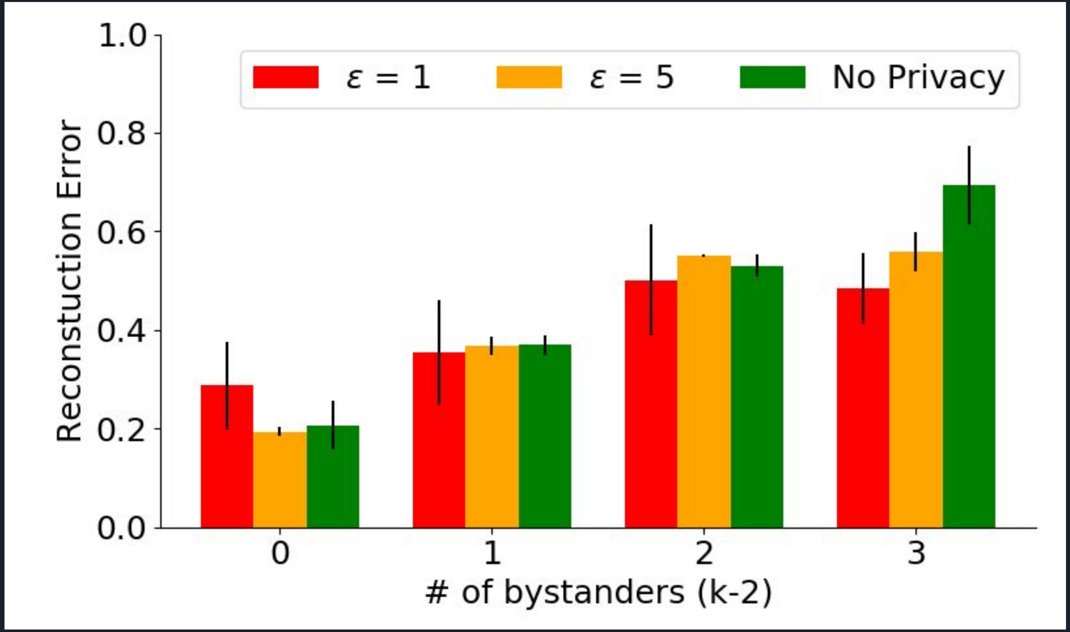


Inversion Attack Defenses

- Prior work: defending model inversion
  - Prevent attacker from **re-creating the victim's model**
  - Mitigated by privacy parameters (all specified by client)
    - Setting a stronger privacy parameter ε
    - More "bystanders" in model training

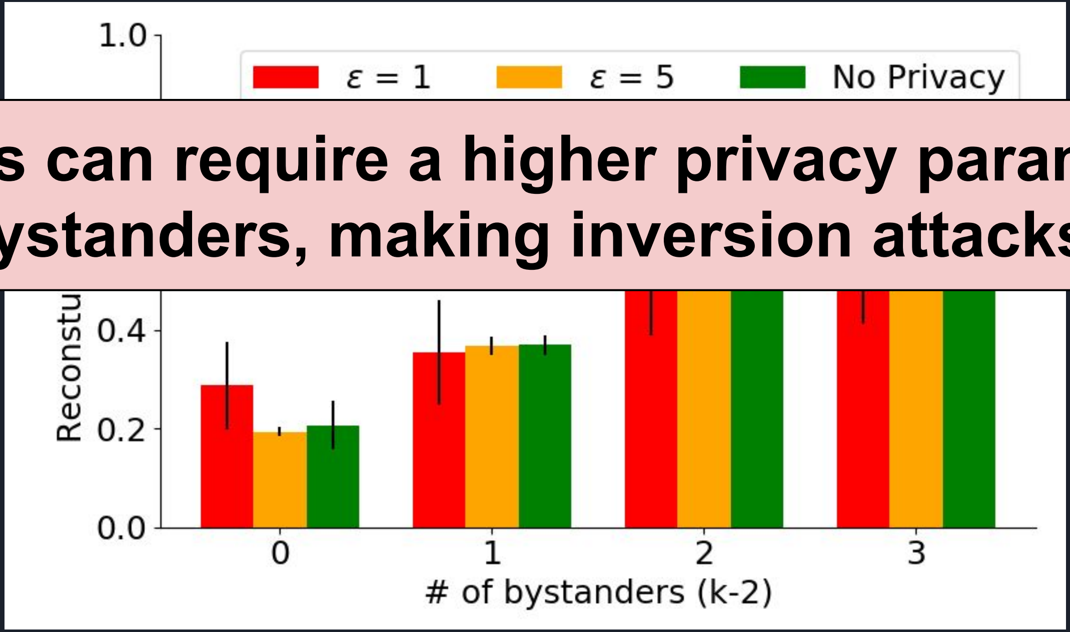
Evaluation: Inversion Defense

- Run a TorMentor inversion attack with increasing bystanders



Evaluation: Inversion Defense

- Run a TorMentor inversion attack with increasing bystanders



Clients can require a higher privacy parameter or more bystanders, making inversion attacks weaker.

Poisoning Attack Defenses

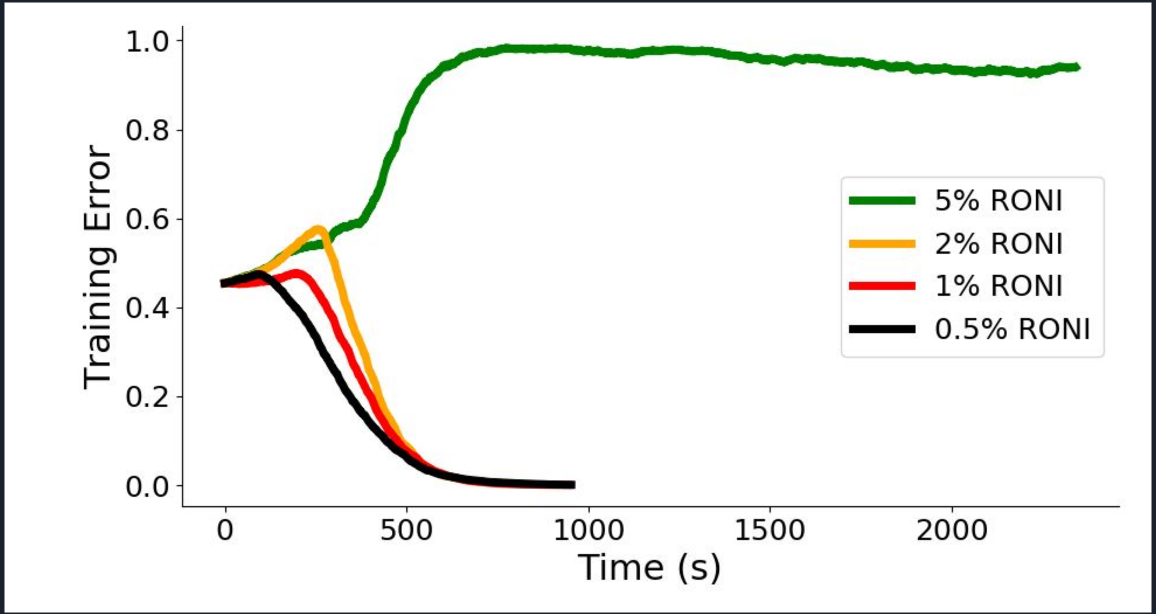
- Reject on Negative Influence (RONI)
  - Reject datasets with negative impact on “influence” metric

- We implemented a distributed RONI:
  - Evaluate influence of model updates instead of data
  - Use curator provided validation set
  - Reject clients that exceed defined threshold

[1] Nelson et al. “*Exploiting machine learning to subvert your spam filter*” LEET '08

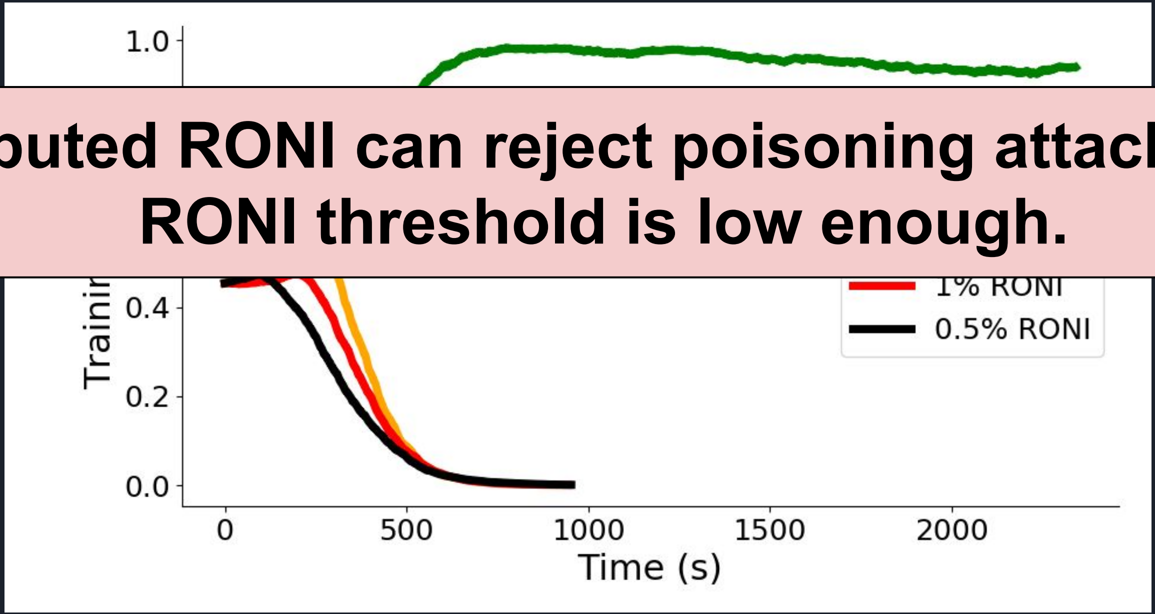
Evaluation: Poisoning Defense

- Test varying RONI thresholds against a system of 75% attackers



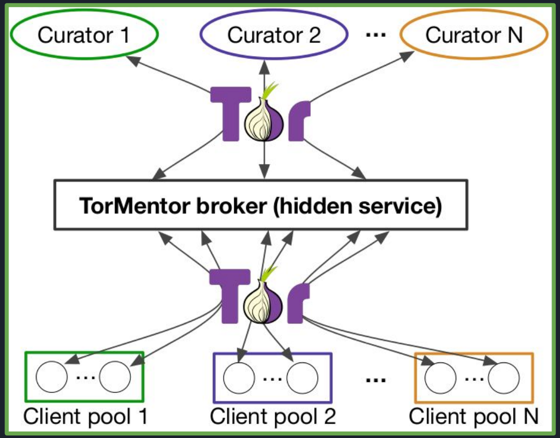
Evaluation: Poisoning Defense

- Test varying RONI thresholds against a system of 75% attackers



Distributed RONI can reject poisoning attacks if the RONI threshold is low enough.

TorMentor contributions



- Existing ML systems do not provide:
  - Anonymity, privacy

- We propose **brokered learning** to facilitate anonymous and privacy-preserving ML

- TorMentor prototype
  - Clients and curators are unaware of each other
  - Supports client churn, heterogeneous privacy parameters
  - Actively monitors and rejects malicious clients

https://github.com/DistributedML/TorML

Bonus

Threat Model

- Guarantees:
  - Broker honours privacy parameters
  - Anonymity, same as Tor
  - Defending Sybils, same as proof of work

- Assume:
  - Adversaries know of and can target clients or broker
  - Sybil attacks possible: curators and clients can collaborate
  - Users adhere to the given APIs for joining and SGD