Securing Federated Sensitive Topic Classification against Poisoning Attacks
Tianyue Chu\textsuperscript{1}, Alvaro Garcia-Recuero\textsuperscript{3}, Costas Iordanou\textsuperscript{3}, Georgios Smaragdakis\textsuperscript{3}, Nikolaos Laoutaris\textsuperscript{3}
\textsuperscript{1}IMDEA Networks Institute \textsuperscript{2}Cyprus University of Technology \textsuperscript{3}TU Delft

Motivation
- Legislation: safeguard online content that contains sensitive data
- Centralized classifier: tied to a fixed training set and cannot be used to drive a privacy-preserving distributed classification system
- A Federated learning (FL) solution: continuously learn from real-time web data gathered by users and can be distributed with privacy

Contributions
- A FL classifier: classify arbitrary URLs that contain GDPR sensitive content, achieving comparable accuracy with the centralized one
- Robust for poisoning attacks: our FL solution combines a reputation score with residual-based attack detection
- EITR: implement our FL classifier in a prototype system (EITR) and validate it with real users

Theoretical Guarantees
The Corollary shows the convergence rate and error rate

Corollary
Continuing with Theorem 1, when the iterations satisfy
\[
t \geq \frac{1}{\epsilon} \log \left( \frac{1}{3\epsilon^2} \right), 3\epsilon \in \left[ 0, \frac{4d}{1+4M\epsilon^2} \right], \] we have:
\[
P\left( \left| w^t - w^\star \right| \leq \frac{2\sqrt{N}}{L} \Delta_t + \frac{2}{L} \Delta_0 \right) \geq 1 - \xi
\]
- The convergence is guaranteed in bounded time
- The trade-off between convergence rate and error rate
- Guidance for hyper-parameters tuning

Performance Evaluation
- Figure 4: Training Loss and Average Accuracy for seven evaluated methods
- Figure 5: The change of performance metrics for varying percentage of attackers for seven evaluated methods

Real-user Experiment
We had 50 users participating in our experiment using EITR
- The majority of users have reputation scores falling in the intermediate range
- Our method converges as rapidly as in simulation and achieves an average accuracy of 80.36%
- The ROC curve in real-user experiment yielded 0.79 AUC

Broader Impact
- NDSS '23: Our paper is accepted by the Network and Distributed System Security Symposium 2023 - Summer Review Cycle
- The ONE conference 2022: Our team was invited to present our work at ONE conference 2022.