VERS UN APPRENTISSAGE FÉDÉRÉ ROBUSTE FACE AUX ATTAQUES BYZANTINES

L'IA de confiance exige équité, confidentialité et robustesse. Ces défis sont amplifiés dans les systèmes d'apprentissage fédéré du fait de leur nature distribuée. Nous nous appuyons sur le progrès des attaques d'inférence afin d'éliminer les activités Byzantines, sans compromis de confidentialité grâce au chiffrement homomorphe.

Doctorant

Adda Akram BENDOUKHA

Encadrement

Nesrine KAANICHE

Renaud SIRDEY

Aymen BOUDGUIGA

Co-auteurs

Aymen BOUDGUIGA

Nesrine KAANICHE

Renaud SIRDEY

Didem Demirag

Sébastien Gambs

Partenaires

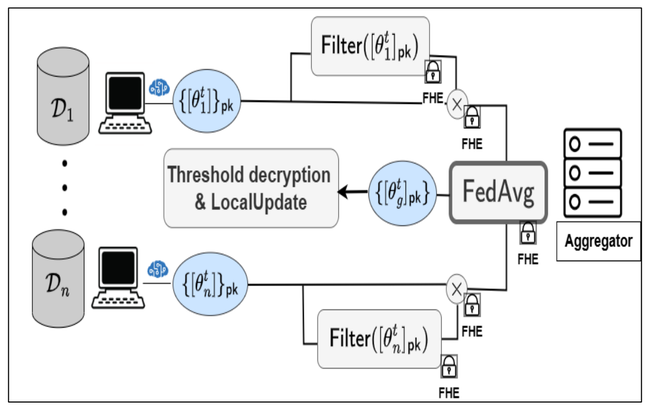
Soutenu par

Projet EQUIHID

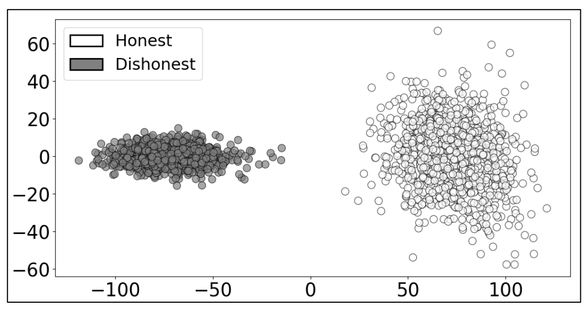
Contexte

La confidentialité des mises à jour des clients en apprentissage fédéré renforce les vulnérabilités face aux attaques byzantines. Nous proposons une approche combinant **chiffrement homomorphe (FHE)** pour la confidentialité et **filtrage par inférence SVM dans le domaine chiffré**. Contributions principales :

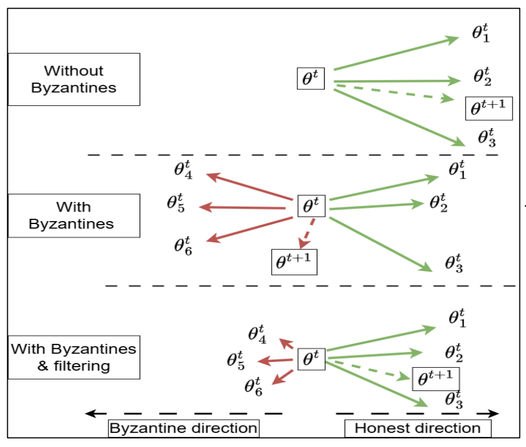
- Détection des attaques Byzantines par inférence des mises-à-jour chiffrées.
- Optimisation du modèle de filtrage pour l'efficacité de l'inférence dans le domaine chiffré (**FHE-friendliness**).
- Conception des **primitives FHE nécessaires à l'inférence de SVM**.
- Évaluation complète de l'efficacité et de la **praticité face à diverses attaques**



FL avec filtrage chiffré des updates



Filtre: Projection PCA+ SVM linéaire



Impact du filtrage sur l'apprentissage collaboratif avec moyenne

Algorithm 3 Filtering data generation for a property P

Require: A base dataset D_{shadow}, number of epochs T
Ensure: Dataset $\mathcal{D}_P = \{(\Delta_i, y_i)\}_{i=1}^{K}$ for a property P

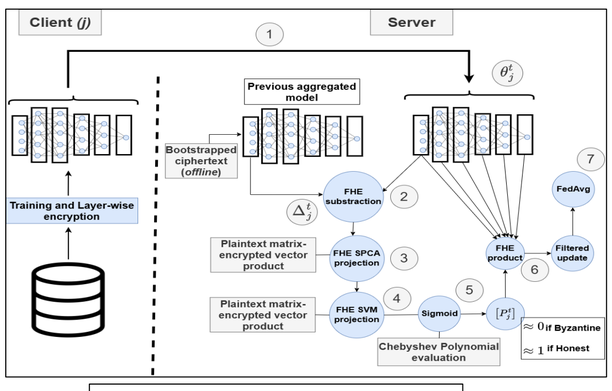
$\mathcal{D}_P \leftarrow \emptyset$ ▷ Initialize dataset
for $i = 1$ to R **do**
  $\theta^{\text{init}} \leftarrow \text{RandomInit}(\text{Seed}_i)$
  $D^i_{\text{shadow}} \leftarrow \text{DirichletSample}(D_{\text{shadow}}, \alpha = 1.0)$
  for $t = 1$ to T **do**
    $\theta^{t+1} \leftarrow \text{HonestUpdate}(\theta^t, D^i_{\text{shadow}})$
    $\bar{\theta}^{t+1} \leftarrow \text{DishonestUpdate}_P(\theta^t, D^i_{\text{shadow}})$
    $\Delta_t = \theta^{t+1} - \theta^t$ ▷ Honest update difference
    $\bar{\Delta}_t = \bar{\theta}^{t+1} - \theta^t$ ▷ Byzantine update difference
    $\mathcal{D}_P \leftarrow \mathcal{D}_P \cup (\Delta_t, 1)$ ▷ Label: 1
    $\mathcal{D}_P \leftarrow \mathcal{D}_P \cup (\bar{\Delta}_t, 0)$ ▷ Label: 0
  end for
end for
return $\mathcal{D}_P$

Génération de updates honnêtes et Byzantines pour plusieurs attaques, et constitution d'une base de données de ***shadow updates*.**

Méthodologie:

Nous adaptons les **attaques d'inférence de propriété** pour identifier les comportements byzantins.

- Un modèle de Machine Learning, entraîné sur des **shadow models**, apprend à reconnaître les signatures d'attaques *(backdoor, gradient ascent, label flipping)*.
- Utilisation d'un **SVM linéaire optimisé pour l'inférence en FHE**.
- Compatibilité avec l'agrégation sécurisée grâce au **faible coût d'inférence des SVM linéaires et polynomiaux** dans le domaine chiffré



Processus du filtrage en plusieurs étapes

Dataset	Model Architecture	Attack Type
FEMNIST	LeNet-5 [79]	Label shuffling
CIFAR-10	ResNet-14 [24]	Gradient Ascent
GTSRB	VGG11 [25]	Backdoor
ACSIncome	3-layer MLP	Label Flipping

Données, architectures de modèles, et types d'attaques expérimentés

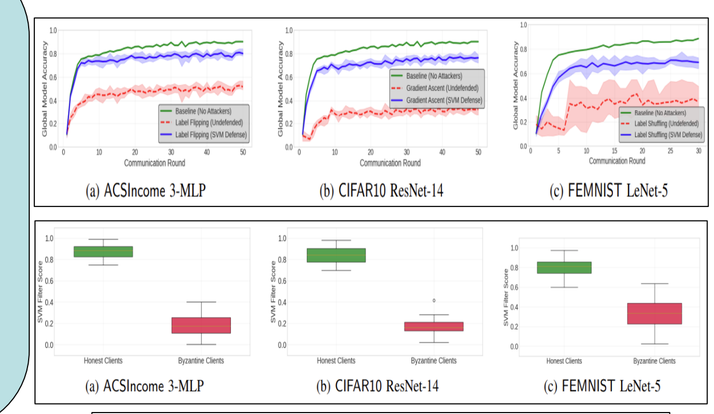
Résultats & Perspectives

Résultats :

- Convergence maintenue avec **jusqu'à 50 % de clients byzantins**.
- **Temps d'inférence chiffrée < 7 s**, démontrant la praticité de l'approche.

Perspectives :

- Étudier la détection de **multiples comportements byzantins** via un **unique SVM**.
- Évaluer la robustesse lorsque les mises à jour sont générées par **DP-SGD** (confidentialité différentielle).



Trajectoires de convergences et distributions des valeurs de filtres

PAPIER (ARXIV): ROBUST FEDERATED LEARNING VIA BYZANTINE FILTERING OVER ENCRYPTED UPDATES

Contact adda-akram.bendoukha@telecom-sudparis.eu

@equihid

RST PhD Day

Février 2026