3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Ph.D. research topic

- Title of the proposed topic: Attacks and Defenses against Federated Learning
- Research axis of the 3IA: 4
- **Supervisor (name, affiliation, email): Melek Önen, EURECOM, melek.onen@eurecom.fr**
- Potential co-supervisor (name, affiliation): Ayşe Ünsal, EURECOM
- The laboratory and/or research group: EURECOM

Apply by sending an email directly to the supervisor.
The application will include:

- Letter of recommendation of the supervisor indicated above
- Curriculum vitæ.
- Motivation Letter.
- Academic transcripts of a master's degree(s) or equivalent.
- At least, one letter of recommendation.
- Internship report, if possible.

⇨ **All the requested documents must be gathered and concatenated in a single PDF file named in the following format: LAST NAME of the candidate_Last Name of the supervisor_2023.pdf**

- Description of the topic:

Federated learning (FL) enables multiple stakeholders to collaboratively train robust machine learning (ML) algorithms without exchanging the actual data. The benefits of such a decentralized technology over personal and confidential data are multiple and already include some initial privacy guarantee over the underlying data since these never leave the local site. However, several security and privacy issues remain unsolved. Indeed, several research works [1] show that such technology suffers from various privacy leaks even when the data do not leave the local site. Attacks range from (i) inference attacks that aim at extracting some private information about the training data, its features, or the actual model parameters, to (ii) poisoning attacks that target the integrity of the ML model and cause a general degradation of the performance of the model, or the misinterpretation of some inputs to result in a particular behaviour that is favourable to the attacker.

In this thesis, the PhD candidate will first study the impact of decentralizing ML algorithms on their vulnerabilities against those attacks. While, the existing recent literature on the study of such attacks for FL mostly concentrates on deep learning. The PhD candidate will also investigate different ML algorithms such as principal component analysis (PCA) [2] as well as new types of attacks like link stealing attacks [3] whereby the protected information is not just a dataset but has more complex structure (such as graph neural networks). Finally, as initiated

in [4], a stronger adversary model can also be defined and studied whereby the active adversary who wishes to harm the model is aware of the underlying privacy protection mechanism and uses it as a weapon to avoid being detected.

Further to the analysis and identification of potential attacks against FL, the goal of the PhD is to propose adequate defense strategies. These defense strategies range from solutions based on the notion of differential privacy and/or alternative privacy models. The vast majority of DP-FL schemes rely on deep neural networks (NN) models. Little attention has been paid to other models like graph NNs (GNNs) or PCA. Among the few existing works in the literature, [2] proposes an FL algorithm to compute PCA in a DP fashion, but the authors evaluate it just for extremely weak configurations of the privacy parameters. The candidate will aim to develop DP-FL schemes based on GNN and PCA [5] which, on the one hand, can provide acceptable privacy-utility trade-offs for sensible choices of the privacy parameters, and on the other hand, those schemes can mitigate a selection of the attacks identified throughout this study.

Finally, while DP and its variants can indeed be deployed in FL as potential defense mechanisms, the goal of these solutions is to increase the accuracy as much as possible and hence they inherently come at the cost of more privacy leakage [6]. Studies typically conduct empirical evaluations of the actual privacy provided via inference attacks. This detracts from DP as a privacy model, since the main advantage of a privacy model is precisely setting the level of privacy by design, without having to assess it empirically. Accordingly, the PhD candidate will explore the application or adaptation of alternative privacy models and defense mechanisms to DP.

Bibliography

[1] P. Kairouz, H. McMahan, B. Avent, A. Bellet and e. al., "Advances and Open Problems in Federated Learning," *New Foundations and Trends,* 2021.

[2] A. Grammenos, R. Mendoza Smith, J. Crowcroft and C. Mascolo, "Federated principal component analysis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[3] X. He, J. Jia, M. Backes, N. Z. Gong and Y. Zhang, "Stealing Links from Graph Neural Networks," in *Usenix Security Symposium*, 2021.

[4] A. Ünsal and M. Önen, "A statistical threshold for adversarial classification in laplace mechanisms," in *IEEE Information Theory Workshop (ITW)*, 2021.

[5] O. Zari, J. Parra-Arnau, A. Ünsal, T. Strufe and M. Önen, "Membership inference attack against principal component analysis," in *Proc. Privacy in Statistical Databases (PSD)*, 2022.

[6] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *USENIX Security Symposium*, 2019.