Project title: **Enhancing Privacy-Preserving Federated Learning with Incentives**

Contact. Primary Supervisor: Dr Aydin Abadi, email address: aydin.abadi@ncl.ac.uk
Co-supervisor: Dr Mohammad Naseri (Flower Labs).

Research project

Abstract. This research aims to design a secure, reward-based mechanism that incentivizes data contributors to provide their sensitive input data to a collaborative, privacy-preserving federated learning process while ensuring the integrity and privacy of the process. This has the potential to yield a fairer system and advance the goals of **ethical machine learning** and **responsible AI**. By leveraging techniques from federated learning, cryptography, and game theory, the project will develop technical solutions that fairly reward participants based on their contributions and protect against malicious actors. The expected outcomes include a secure formal model, provably secure protocols, practical implementation, and real-world applications in sectors like healthcare and finance, ultimately enhancing the real-world adoption of Federated Learning.

1. Context. The application of machine learning (ML) has been rapidly expanding among different sectors, such as healthcare, finance, and law enforcement. To enhance the quality of model training during ML, it is essential to collect data from different parties that collaborate to train "models." Privacy concerns often arise among these parties. Federated Learning (FL) [1] addresses this by enabling collaborative model training without sharing sensitive data.[1] It involves training a "global model" via collaborative learning on local data and transmitting only model updates, not raw data, to a central server. Overall, an FL includes a server and a set of clients who contribute sensitive data. FL has a wide range of applications, such as dealing with financial fraud [2], disease detection [3], smart city [4], and edge computing [5]. FL and its variants are not merely theoretical concepts; FL start-ups and companies actively implement them. Examples include Flower Labs [6], Sherpa.ai [7], NVDIA [19], and IBM Corporation [8].

2. Critical Limitations of State-of-the-art. Research on enhancing the security and efficiency of privacy-preserving machine learning, especially FL, is evolving at a fast pace. There are three main facts about any secure privacy-preserving computation mechanism (such as FL). Firstly, a secure privacy-preserving mechanism output always reveals some information about the sensitive input data [20]. For instance, the final global model developed in FL reveals some information about each party's input. Secondly, not everyone contributing their private input is necessarily interested in the result. Thirdly, participating in a privacy-preserving mechanism, like FL, imposes computation and computation costs on the participants. Most FL schemes have assumed that parties participate in the FL process free of charge, bearing the burden of privacy, computation, and communication overheads. The importance of incentivizing parties in FL has been recognized, e.g., see [9, p. 15]. There have been efforts to incentivize participants of FL [10,11,12, 13, 14]. However, these schemes assume that the parties will follow the procedure's instructions and are paid after the procedure is fully completed. Nevertheless, the existing schemes will not be secure if some of the parties are malicious and deviate from the schemes' instructions. Malicious parties can act in their favour. Malicious parties can abort prematurely during the execution of FL and learn other parties' sensitive data without allowing others to learn the result, e.g., the global model. To date, there is no incentive mechanism for FL in the literature that resists malicious adversaries.

3. Research Question and Hypothesis. The project will design novel solutions to mitigate FL's incentivization issue, addressing the question:

> Main question: *How can we devise a provably secure mechanism that rewards contributors of FL, while resisting malicious actors?*

This question is vital for the following reasons: (a) As FL's utility across various sectors like healthcare and finance is becoming increasingly apparent, ensuring security and incentivizing contributors becomes paramount. (b) FL confronts threats like influence on result integrity and sensitive data inference. A secure mechanism is crucial to identify and mitigate these threats. (c) Rational contributors in FL require incentives to participate in model training. A robust reward

[1] All citations in this document are hyperlinked.

mechanism can incentivize participation and collaboration in FL, hence enhancing real-world FL adoption. (d) Establishing trust among participants is crucial for FL initiatives to succeed. A provably secure mechanism not only guarantees the integrity of the FL process but also enhances transparency, enabling participants to have confidence in the system's reliability. The research's objective is to significantly improve FL and facilitate its real-world adoption by devising an effective mechanism for rewarding FL data contributors. It must ensure security, integrity, and privacy in the face of evolving threats. To realize this objective, the research will rely on the following hypothesis.

> Main Hypothesis: *Providing financial rewards that are proportionate to the extent of participants' contributions can effectively incentivize their involvement*.

To validate the hypothesis, the research will involve three complementary phases:

- Phase 1: **Formal modeling**. To establish a scientific foundation for fundamental security assurances required to incentivize FL participants. The objective is to formalize security measures to encourage participation in the FL procedure and deter malicious actors from compromising result accuracy or manipulating reward distributions.
- Phase 2: **Development of provably secure security protocols**. Phase 2's objective is to develop a secure protocol that aligns with the model, adequately compensating participants in FL proportionate to the amount of information they contribute to the procedure.
- Phase 3: **Implementation and evaluation**. In this phase, the protocols developed in Phase 2 will be implemented (in Python) for evaluation and establishing concrete parameters. The research will collaborate with the software development team in Flower Labs to ensure the scalability of the prototype.

4. The Project's Outcomes and Impacts. The outcome of each phase is as follows: (1) This will be the first mathematical model for FL mechanisms required to incentivize participants and resist malicious actors. This model will establish a foundation for systematically assessing other FL variants. The outcome will also benefit the research community in information security, FL, and cryptography; (2) It will be the first mechanism that will simultaneously achieve two key objectives: firstly, enabling FL participants to receive rewards, and secondly, ensuring security even in the presence of malicious actors; and (3) It will be the first open-source package that will implement the protocols developed in Phase 2. The implementation and benchmark also represent a contribution to the FL and cryptography research communities, as they will provide a basis for building other related protocols in the future. The implementation has the potential to be integrated into the Flower Labs software.

5. The Project's Societal Importance and Impact. FL has many applications, from combatting financial fraud [2] to detecting online grooming [15] and enhancing healthcare services [16]. This research seeks to enhance FL's real-world adoption. By increasing FL's adoption, UK residents will benefit through improved services in the healthcare, finance, and retail sectors. FL aids in detecting and addressing online grooming. Thus, it can lower the emotional distress and trauma associated with such incidents while preserving individuals' privacy.

6. Project Timeline. (a) Months 1—13: Conduct a comprehensive literature review and acquire the necessary foundational knowledge, (b) Months 12—24: Complete Phase 1 and draft the first paper (a survey paper), (c) Months 23—36: Complete Phase 2 and write the second paper (Systematization of Knowledge), (d) Months 32—42: Complete Phase 3 and prepare the third paper (describing the fully developed system), and (e) Months 41—48: Finalize the thesis and write the fourth paper (focused on improving the efficiency of the solution presented in the third paper).

7. Supervision Environment. This PhD project has been codesigned and will be jointly supervised by Dr. Aydin Abadi, an expert in cryptography and privacy-preserving machine learning, and Dr. Mohammad Naseri [18] from Flower Labs, a leading organization in federated learning solutions [6]. This collaboration provides the candidate with a combination of academic excellence and industry-driven innovation, offering exposure to state-of-the-art research and practical challenges in federated learning.

8. Applicant skills/background. Candidates should possess or be highly motivated to acquire, strong knowledge in the following areas: (1) cryptography, (2) Privacy-enhancing technologies, such as FL, (3) mathematics (including number theory and game theory), and (4) computer programming in C++, Java, or Python.