Ph.D.: Distributed Machine Learning in Ubiquitous Environments using Location-dependent Models

June 2022

1 Contact Information and Supervisory Team

Supervisor Team:	Romain Rouvoy (PU, U. Lille), Lionel Seinturier (PU, U. Lille) Davide Frey (CR, Inria), François Taiani (PU, UR1) {romain.rouvoy,lionel.seinturier}@univ-lille1.fr, {davide.frey,francois.tainai}@irisa.fr
Laboratories:	Inria Centre at the University of Lille & Inria Centre at Rennes University
Research Group:	WIDE (the World Is DistributEd) and SPIRALS (Self-adaptation for distributed services and large software systems)
Location:	Lille or Rennes (France)
Funding:	three-year fixed-term work contract (with standard French health-benefits and retirement plan)
Keywords:	Machine Learning, Decentralization, Privacy Protection, Middleware, Crowd-sensing, Geolocation

2 Context and Motivation

Mobile devices are producing a deluge of data by leveraging a wide variety of embedded or connected sensors that capture the surrounding environment of end-users and their routines. However, this continuous data stream inevitably includes sensitive information that may jeopardize the privacy of end-users if processed by malicious stakeholders. **While machine learning algorithms are nowadays widely adopted as a convenient keystone to process large datasets and infer actionable insights, they often require grouping the raw input data into a single place, thus imposing a privacy threat for end-users sharing their data.**

To address this ethical challenge, *privacy-preserving machine learning* [16] and *decentralized machine learning* (DML) [2, 17] are revisiting state-of-the-art machine learning algorithms to enforce user privacy, among other properties. While most of the research contributions that have been achieved in this area remain at the theoretical level [15], the implementation and the deployment of these privacy-preserving algorithms in the field remain a challenging issue for most modern applications, at best resulting in *ad hoc* research prototypes [3].

Among difficulties, the limited resources of mobile devices and their partial *device-to-device* (D2D) connectivity makes it challenging to adopt DML algorithms for the masses. Nonetheless, we claim that leveraging fleets of mobile devices may provide promising opportunities for DML algorithms in the context of mobile crowdsourcing software systems [7]. In particular, we believe that unsupervised machine learning algorithms (such as clustering), and federated learning algorithms, can benefit from nearby devices to reason upon a reduced set of relevant samples, captured as models by nearby devices sharing similar concerns and objectives. As promoted by DML approaches, the aggregation of such *in situ* models can effectively contribute to delivering **personalized results (e.g. recommendations) to end-users without exposing their privacy**.

3 Ph.D. Objectives

In many applications, machine learning models are intrinsically tailored to a given geographical area. This is the case, for example, of smart building management [4], crowdsensing for environmental monitoring [9], and smart wireless transmission techniques [10]. These models benefit from continuous data streams generated by sensors and/or mobile devices. **The goal of this Ph.D. is to design, deploy and characterize decentralized learning algorithms and frameworks that can preserve the privacy of their users, while delivering location-dependent services.**

3.1 Research Questions & Work Plan

More specifically, our objective in this PhD project is to investigate **how decentralized machine learning can be effectively deployed on mobile devices to design and implement location-dependent applications accessible to end-users in the field, while preserving their privacy**. This objective calls for a combination of novel algorithms, protocols, and middleware solutions. More specifically, we foresee that achieving this objective requires addressing three challenges:

1. **How to store unbounded data streams on constrained mobile devices?** DML enforces the local processing of data but depending on sensors, the volume of produced data streams may quickly go beyond the storage capacity of mobile devices. We, therefore, intend to leverage our past work on temporal graph storage techniques [6] to store compact representations of data streams and support reasoning over long histories of sensor data. **Intended duration:** 12 months;

2. **How to exchange relevant model samples among nearby devices?** By connecting nearby devices, DML algorithms can periodically exchange locally learned models and knowledge. Yet, one cannot assume to blindly share these local models for privacy concerns. Instead, the exchange of partial models of shared interests should be privileged. This requires the design of a new privacy-preserving protocol to identify common interests (*e.g.*, shared locations) and then extract the associated partial model that can be shared among connected parties. To this end, we plan to leverage our work on privacy-preserving decentralized averaging [11, 1, 3], in order to compute aggregate gradients without releasing sensitive data. We also plan to build on recent work on the convergence of stochastic gradient methods in the presence of data streams generated by Markov processes [5, 14] (note that these papers ignore space correlations). **Intended duration:** 12 months;

3. **How to program DML algorithms for the masses?** Finally, to avoid the design and implementation of *ad hoc* algorithms, we aim to design a middleware framework that can support the execution of a larger family of DML algorithms. This programming framework will nonetheless include a mobile testing environment supporting nearby communications in order to tune the hyper-parameters of such models [8]. **Intended duration:** 12 months.

As a matter of demonstration and assessment of the contributions to the above challenges, we intend to consider two case studies in the area of mobile crowdsourcing software systems:

- *Spatio-temporal clustering of air quality measurements* will aim to support a decentralized inference of a particle matters cartography by leveraging a network of air quality sensors connected to personal smartphones. While *in situ* measurements may disclose points of interest, we aim at computing custom maps of particle matter concentrations in order to recommend end-users appropriate locations with low pollution;

- *User-centric clustering of user trajectories* will consider the more general case of individual trajectory processing to extract shared paths explored by a crowd of users, and possibly predict future user's mobility. These shared paths and trajectory predictions can in turn be used for task planning (in a crowdsensing scenario) and resource allocation (for instance, in a multi-access edge computing (MEC) framework). However, user trajectories cannot be shared without disclosing routines or sensitive places. Thus we aim to build on the vicinity of end-users in order to infer anonymized mobility models. A particular approach we will explore is to build on the vicinity of users, via standard localization methods or channel charting based solely on radio measurements from personal smartphones [13], in order to infer anonymized mobility models.

The above crowdsensing case studies will draw upon the expertise gathered within the SPIRALS team on crowdsensing platforms, particularly through the APISENSE [9] online platform. The Ph.D. might also involve experiments on specialized testbeds, such as FIT IoT-LAB, which can be used to collect IoT data and experiment with actual ML and FL implementations.

4 Profile of the Optimal Candidate for this PhD

The candidate recruited for this Ph.D. should have a Master's Degree in Computer Science or equivalent, with **a solid algorithmic and systems background**, particularly regarding at least one of the following: distributed computer systems, machine learning, and/or mobile computing. Good programming skills and **a willingness to learn about new techniques** (decentralized machine learning and privacy protection) are also crucial, as well as good writing skills and the ability to propose, present, and discuss new ideas in a collaborative setting.

5 The Fed-Malin project

The proposed Ph.D. will take place within the **Fed-Malin Inria Challenge project**. FED-MALIN aims to address the methodological challenges of moving ML operations from the comfortable cloud nest to the wild Internet. Most existing research considers the "Google/Apple setting" with a large set of relatively homogeneous smartphones and the cloud. By contrast, we will consider various scenarios, including entities with significant computation resources (e.g., companies, hospitals), edge servers deployed by telecommunications operators, and (potentially heterogeneous) IoT devices with or without AI edge accelerators.

FED-MALIN will shed light on the design principles of distributed systems for ML. It will help to better configure the existing ones, as well as to conceive the next generation. More efficient distributed learning systems can reduce cost barriers to access AI technology and mitigate the present concentration of power at the few technology giants that can afford massive computing power. They may also operate under lower energy budgets and thus potentially contribute to making AI more sustainable [12].

References

[1] Tristan Allard, Davide Frey, George Giakkoupis, and Julien Lepiller. Lightweight privacy-preserving averaging for the internet of things. In *Proceedings of the 3rd Workshop on Middleware for Context-Aware Applications in the IoT, M4IoT@Middleware 2016, Trento, Italy, December 12-13, 2016*, pages 19–22. ACM, 2016.

[2] Aurélien Bellet, Rachid Guerraoui, Mahsa Taziki, and Marc Tommasi. Fast and differentially private algorithms for decentralized collaborative machine learning. *CoRR*, abs/1705.08435, 2017.

[3] Pierre Dellenbach, Aurélien Bellet, and Jan Ramon. Hiding in the crowd: A massively distributed algorithm for private averaging with malicious adversaries. *CoRR*, abs/1803.09984, 2018.

[4] Djamel Djenouri, Roufaida Laidi, Youcef Djenouri, and Ilangko Balasingham. Machine learning for smart building applications: Review and taxonomy. *ACM Computing Surveys (CSUR)*, 52(2):1–36, 2019.

[5] Thinh T. Doan, Lam M. Nguyen, Nhan H. Pham, and Justin Romberg. Finite-time analysis of stochastic gradient descent under markov randomness. *CoRR*, abs/2003.10973, 2020.

[6] Thomas Hartmann, François Fouquet, Assaad Moawad, Romain Rouvoy, and Yves Le Traon. Greycat: Efficient what-if analytics for data in motion at scale. *Inf. Syst.*, 83:101–117, 2019.

[7] Lakhdar Meftah, Romain Rouvoy, and Isabelle Chrisment. FOUGERE: user-centric location privacy in mobile crowdsourcing apps. In *DAIS*, volume 11534 of *Lecture Notes in Computer Science*, pages 116–132. Springer, 2019.

[8] Lakhdar Meftah, Romain Rouvoy, and Isabelle Chrisment. Testing nearby peer-to-peer mobile apps at large. In *MOBILESoft@ICSE*. IEEE / ACM, 2019.

[9] Lakhdar Meftah, Romain Rouvoy, and Isabelle Chrisment. Empowering mobile crowdsourcing apps with user privacy control. *J. Parallel Distributed Comput.*, 147, 2021.

[10] Timothy O'Shea and Jakob Hoydis. An introduction to deep learning for the physical layer. *IEEE Transactions on Cognitive Communications and Networking*, 3(4):563–575, 2017.

[11] Amaury Bouchra Pilet, Davide Frey, and François Taïani. Robust privacy-preserving gossip averaging. In Mohsen Ghaffari, Mikhail Nesterenko, Sébastien Tixeuil, Sara Tucci, and Yukiko Yamauchi, editors, *Stabilization, Safety, and Security of Distributed Systems - 21st International Symposium, SSS 2019, Pisa, Italy, October 22-25, 2019, Proceedings*, volume 11914 of *Lecture Notes in Computer Science*, pages 38–52. Springer, 2019.

[12] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696, Apr. 2020.

[13] Christoph Studer, SaïD Medjkouh, Emre Gonultaş, Tom Goldstein, and Olav Tirkkonen. Channel charting: Locating users within the radio environment using channel state information. *IEEE Access*, 6:47682–47698, 2018.

[14] Tao Sun and Dongsheng Li. Decentralized Markov Chain Gradient Descent. *CoRR*, abs/1909.10238, 2021.

[15] Paul Vanhaesebrouck, Aurélien Bellet, and Marc Tommasi. Decentralized collaborative learning of personalized models over networks. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pages 509–517. PMLR, 2017.

[16] Kaihe Xu, Hao Yue, Linke Guo, Yuanxiong Guo, and Yuguang Fang. Privacy-preserving machine learning algorithms for big data systems. In *ICDCS*, pages 318–327. IEEE Computer Society, 2015.

[17] Blaise Agüera y Arcas. Decentralized machine learning. In *BigData*, page 1. IEEE, 2018.