Towards a Federated Learning Approach for Privacy-aware Analysis of Semantically Enriched Mobility Data

Zaineb Chelly Dagdia
Université Paris-Saclay, UVSQ, DAVID
France
zaineb.chelly-dagdia@uvsq.fr

Chiara Renso
ISTI - Institute of National research Council of Italy
Pisa, Italy
chiara.renso@isti.cnr.it

Karine Zeitouni
Université Paris-Saclay, UVSQ, DAVID
France
karine.zeitouni@uvsq.fr

Nazim Agoulmine
Université Paris-Saclay, Univ Evry, IBIS
Evry-Courcouronnes, France
nazim.agoulmine@univ-evry.fr

ABSTRACT

Today, Artificial Intelligence is still facing a major challenge which is the fact of handling and strengthening data privacy. This challenge rises from the collected data which are associated with the fast development of mobile technologies, the huge capacities of high performance computing, and the large-scale storage in the cloud. In this paper, we focus on a possible solution to this challenge which is the use and application of federated learning. Specifically, beyond the federated learning based approaches proposed in different application domains, we mainly focus and discuss a federated learning approach for privacy-aware analysis of semantically enriched mobility data. We introduce the main motivation and opportunities of applying federated learning in mobility data, and highlight the main concepts and basics of our approach by describing our objectives and our approaches' requirements. We, also, describe our workplan that will permit achieving our predefined objectives via the setup of several research questions.

CCS CONCEPTS

• **Information systems** → **Data federation tools**; *Information systems applications*; • **Computer systems organization** → **Cloud computing**; • **Security and privacy** → **Privacy protections**.

KEYWORDS

Mobility, Federated Learning, Privacy

ACM Reference Format:
Zaineb Chelly Dagdia, Chiara Renso, Karine Zeitouni, and Nazim Agoulmine. 2021. Towards a Federated Learning Approach for Privacy-aware Analysis of Semantically Enriched Mobility Data. In *Proceedings of the 1st Workshop on Flexible Resource and Application Management on the Edge (FRAME '21), June 25, 2021, Virtual Event, Sweden.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3452369.3463823

FRAME '21, June 25, 2021, Virtual Event, Sweden

ACM ISBN 978-1-4503-8384-4/21/06…$15.00
https://doi.org/10.1145/3452369.3463823

1 CONTEXT AND MOTIVATION

The pervasiveness of smartphones and various connected sensors and wearables is leading to a wide collection of data including positioning and trajectories, reflecting the user's mobility. This movement data can be enriched with various semantic dimensions, either collected by additional mobile sensors, or derived from external sources in the web by map-matching, e.g., with POIs, or localized events. Mobility data analysis and mining is of paramount interest for several applications such as traffic engineering, urban planning, and environmental studies. Nonetheless, individual trajectories pose risks by being highly sensitive information [8], as they can easily become personally identifiable, even when they are pseudo-anonymized and scrubbed of unique personal information. The use of such data has become more and more restricted. Eventually, the risks are exacerbated in the context of semantically enriched trajectories, that encompass more information on the individual habit, and surrounding context. Therefore, there is a great need to develop new solutions that could allow analyzing mobility and trajectories data while preserving the privacy of individuals. Several recent works have tackled this problem in different settings such as in location based services, aggregate queries, and anonymous trajectory publishing.

The question of individual privacy has become a paramount question today with all the collected data associated with the fast development of mobile technologies, the huge capacities of high performance computing, and the large-scale storage in the cloud. In the market, companies propose applications to citizens who install these without (in the general case) knowing exactly what kind of data is collected and how it is really used [17] since they see the short-term benefit of the applications.

While the above-mentioned concerns were not a huge problem in the beginning of the spread of mobile technologies, with the fast development of intelligent and mining techniques in the cloud and the large deployment of mobiles devices, companies are now not only able to collect numerous data from individuals' mobile devices but also to correlate them with data from other sources and eventually infer additional personal information about individuals that go beyond the services being provided [6, 20].

Some countries have already tried to regulate this collection of personal data (e.g. with GDPR in Europe). However, this target is difficult to fully achieve due to the unbalanced relation between

IT companies and individuals: many, if not most, of the mobile applications and services are provided for free and the personal data collected from them represent the value on which companies build their revenue by profiling users and targeting their advertisements. In such a context, individuals are confronted with a dilemma whether to use these applications and accept the risk of losing their privacy or not being able to use these applications and be excluded from the associated innovations (even socially sometimes).

Location Privacy-Preserving Mechanisms (LPPM) have been extensively studied in the last decade, and it is still a hop topic. Most approaches assume a trusted proxy server which acts as an anonymizer of the user's location [7]; and adopt k-anonymity, distortion, or obfuscation mechanisms [16]. Other techniques are based on encryption, but remain impractical due to the high computation costs incurred. More recently, Differential Privacy (DP), and its variant Local Differential Privacy, have been developed and successfully applied to data sanitization for statistical purposes [25]. Geo-indistinguishability extends the definition of DP to deal with location privacy, by adding a distance condition from the user's location [1][5]. But when the data distribution is skewed, this method compensates by adding more noise, which compromises the utility. Another limitation is the loss of privacy when the mechanism is used repeatedly for a user's trajectory, due to the depletion of the privacy budget. This is why Apple uses a privacy budget which limits user contributions to two donations per day [9]. Basically, several techniques were designed for the Location Based Services (LBS) context when the user issues a single LBS query. Dedicated solutions specifically address continuous LBS queries [4], and privacy preserving trajectory publishing [24], [13]. In the context of privacy-preserving mobile participatory sensing, a mobile distributed architecture based on secure mobile devices (e.g., equipped with Intel SGX)[19] has been proposed. It allows accurate computation of spatial aggregates in real-time. In addition to data publishing and statistical queries, there is a great interest by the community in the application of data mining and machine learning on mobility data [27]. Machine learning is already being used in privacy like clustering to adapt the obfuscation, but recent work shows their convergence [15]. Only recently, Federated learning is becoming popular, and few works exist that apply this paradigm on mobility data, except maybe [21], and [11]. Nonetheless, while privacy in semantically rich trajectories has been addressed in [12, 18], incorporating trajectory semantics in differential privacy and/or in federated learning will need further investigation.

The rest of the paper is structured as follows: Section 2 describes the background and the opportunity of Federated Learning. Section 3 presents the objectives and requirements to the problem of inference of mobility trajectory that are preserving by design the privacy of the end-users. Section 4 presents the methodology that will be adopted to achieve the privacy goal. Finally, Section 5 presents the discussion and the main conclusions.

2 BACKGROUND AND OPPORTUNITIES OF FEDERATED LEARNING

Artificial intelligence is taking the stage as the most promising technology for analysing and making predictions over large amounts of data. Machine learning, and specifically deep learning, are the most promising technologies that are receiving huge interest in the literature and as applications in the industries. However, in many applications' fields, the privacy of the data to be analysed presents a high priority. Federated Learning (FL) is a kind of encrypted distributed machine learning technology, in which participants can build a model without disclosing the underlying data, so that the self-owned data does not leave the local device. It is, therefore, a promising solution for developing privacy preserving data analysis for mobility data.

Mobility data, representing the movement of individuals, is not only sensitive from a privacy point of view, but is also a highly complex kind of data as it covers space, time and semantics. These are all aspects that have to be collected, stored and considered for analysis. Despite being a promising solution, federated learning has not been consistently used in mobility data analysis. This is mainly due to the high complexity and heterogeneity of the data that have to be combined with the current challenges of federated learning like the communication costs and the heterogeneity of the environment.

Federated learning has emerged as a new paradigm, where learning algorithms are executed and embedded in mobile clients, such as mobile devices, to collaboratively train a model without exchanging any training data and keeping it solely computed in the devices [14, 22, 26]. This so-called cross-device FL basically applies on a horizontal data partitioning among possibly a large number of devices. Within this schema, the training task is fully distributed and is iterated over the overall or a subset of devices. The global model(s) is (are), therefore, aggregated in a distributed manner. A central server eventually orchestrates the overall process ranging from the installation of the learning algorithms in the mobile devices, to handling their distributed training and convergence as well as the execution of the inference requests without breaking at any point of time the privacy of the end-users.

Our ultimate aim is to investigate, study, and apply this FL approach in the context of privacy preserving mobility mining, which to the best of our knowledge has not been fully investigated yet. Performing such a thorough study is of high importance as mobility mining is considered to be a very specific problem due to its continuous evolution in both time and space. Thus, rethinking the FL algorithms and the overall orchestration architecture is highly required [3, 10].

Another important aspect to mention is when personal mobile devices run analysis algorithms on the edge. In this case, the privacy conditions might dynamically change depending on where these devices are located and if there are other individuals around. It is, therefore, fundamental that such privacy preserving algorithms dynamically recognize the privacy needs and conditions.

Figure 1 shows an abstract view of the architecture where local training and model updates are made at the level of Mobile Clients, and the Federated Learning Server coordinates the participants and the whole FL plan, receives the model updates in multiple rounds and, once the process reaches the stop criteria, the final model is deployed at client side. The use case here is related to privacy-preserving traffic monitoring, but the functioning principle remains the same in other scenarios.

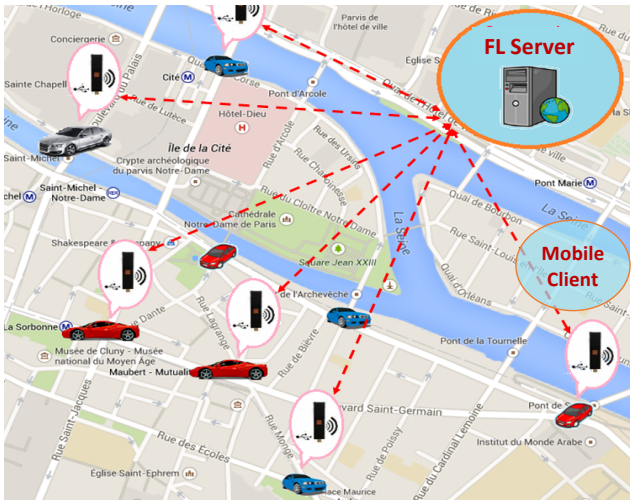

Figure 1: An Abstract View of the FL Approach.

3 OBJECTIVES AND REQUIREMENTS

We envisage the need for novel solutions to the problem of inference of mobility trajectory that are preserving by design the privacy of the end-users. We aim to consider in this work non-functional requirements such as convergence time, accuracy, complexity (of request), etc.

To achieve this objective FL appears as a great solution that permits both avoiding the collected data to be sent outside the mobile devices (privacy preservation) while permitting applications with a distributed logic to extract global parameters from individuals' personal data without necessarily processing them outside the individuals' computers.

Our objective is to propose appropriate solutions for privacy preserving mobility data mining, based on the paradigm of FL. This will consider main mining tasks such as trajectory clustering, mobility pattern mining and mobility prediction.

To make our objective achievable and our FL-based solutions operational, we address three research questions:

- How to efficiently and dynamically decentralize the mining process of mobility data by means of Federated Learning mechanisms?
- How does this solution compare with the conventional centralized setup in terms of model performance and cost efficiency?
- How do we embed privacy-by-design mechanisms in the federated learning tasks?

These general questions can be broken up into several specific challenges and issues that need to be addressed:

- Unlike the conventional vector data, mobility data is a long evolving time sequence. The training instance is tricky. So, how can we adapt the FL pipeline?
- Geographical vicinity is paramount in grouping, discriminating or comparing mobility patterns. How can this characteristic be handled in FL in a way to save the processing and the communication costs?
- So far, FL builds on two types of actors: distributed clients and an orchestrator central server. Many mobility analysis tasks, such as cluster density, and next-location prediction, are based on a geographically close subset of devices, which can be connected at the fog level [3, 10]. Could the FL process be distributed over the edges-fog-cloud architecture?
- How should the existing algorithms of trajectory mining be adapted to FL? Should they be re-designed from scratch? In particular, should the local model training account for the devices' resource limitation [23]?
- As shown for time series, the on-IID (non-Independent and Identically Distributed) data distribution may lead to shifts between the training and testing data [14]. How to remedy this shift in the context of mobility data?
- Which methods of analysis preserving the confidentiality of mobility data are best suited to the architecture, and how to optimize or adapt them?
- How to deploy the FL privacy preservation process on an edge-fog-cloud architecture?

4 WORKPLAN

We intend to design novel solutions for privacy-preserving mobility data mining based on FL and compare them with baseline solutions under different usage contexts and corresponding datasets.

To achieve this goal, we envision a workplan as reported in the following points and illustrated in Figure 2:

(1) Propose a conceptual framework of the FL solutions for privacy-preserving data mining of mobility data: This can be an extension of an existing open-source framework such as FATE [1];
(2) Propose a classification of mobility datasets based on their level of privacy preservation;
(3) Identify the main metrics for the evaluation of the approach in mobile data mining applications;
(4) Propose several alternative designs for FL privacy preserving algorithms and evaluate their performances using the available datasets and configurations: for each task (mobility prediction, traffic forecasting, event detection, etc.), several algorithms can be investigated. We plan to tackle the specific problems related to various existing mobility data models, starting from sequential check-ins, time-use diary, dense GPS trajectories, and finally, multi-aspect trajectories;
(5) Compare performance measures of the proposed solutions against other state of the art approaches for mobility data mining with and without privacy preservation: by leveraging the performance metrics previously specified, we will assess the loss in FL algorithms compared to their traditional ML counterpart, and highlight the privacy-utility trade-off in such context.

The proposal will be demonstrated on real datasets (use cases) collected from various application contexts which are the following: (i) traffic analysis on public data, (ii) privacy-preserving tourism analysis based on social media recordings, and (iii) study of exposure to environmental risk based on mobile crowdsensing [2].

[1] https://fate.fedai.org

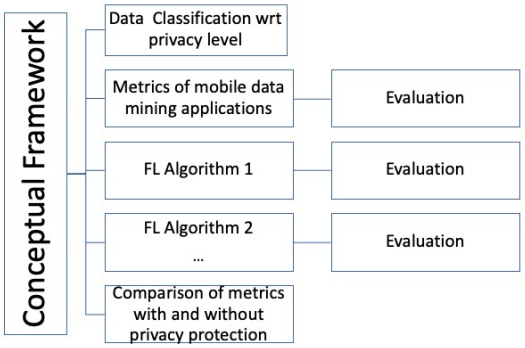

Figure 2: Different steps of the methodology for the evaluation of Federated Learning solutions for privacy preserving on semantically enriched mobility data

5 CONCLUSIONS

In recent years, the emphasis on data privacy has become among the main challenges in artificial intelligence. Federated learning has been considered as a promising solution since it establishes a joint architecture of different connected devices while protecting the local data. In this paper, we have introduced the background and opportunities of federated learning, and discussed its potential in privacy-aware analysis of semantically enriched mobility data. New machine learning methods should be further developed on top of this. Also, new performance metrics of the model quality, the privacy and utility would be necessary to be defined. Yet, it is important to note that applying these metrics for the evaluation of a model in a distributed manner remains a challenge. The federated learning based framework offers plenty of opportunities, and there are many application fields that can benefit from it like, for example, the explosion of the Internet of Things.

ACKNOWLEDGMENTS

This work has been supported by the European Union's Horizon 2020 Research and Innovation program under project MobiDataLab GA101006879.

REFERENCES

[1] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 901–914.

[2] Mariem Brahem, Hafsa E.L. Hafyani, Souheir Mehanna, Karine Zeitouni, Laurent Yeh, Yehia Taher, Zoubida Kedad, Ahmad Ktaish, Mohamed Chachoua, and Cyril Ray. 2021. 12 - Data perspective on environmental mobile crowd sensing. In *Intelligent Environmental Data Monitoring for Pollution Management*, Siddhartha Bhattacharyya, Naba Kumar Mondal, Jan Platos, Václav Snášel, and Pavel Krömer (Eds.). Academic Press, 269–288.

[3] Hung Cao, Monica Wachowicz, Chiara Renso, and Emanuele Carlini. 2019. Analytics Everywhere: Generating Insights From the Internet of Things. *IEEE Access* 7 (2019), 71749–71769. https://doi.org/10.1109/ACCESS.2019.2919514

[4] Yang Cao, Yonghui Xiao, Li Xiong, and Liquan Bai. 2019. PriSTE: from location privacy to spatiotemporal event privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 1606–1609.

[5] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. 2015. Constructing elastic distinguishability metrics for location privacy. *Proc. Priv. Enhancing Technol.* 2015, 2 (2015). https://doi.org/10.1515/popets-2015-0023

[6] Thiago Moreira da Costa, Hervé Martin, and Nazim Agoulmine. 2015. Privacy-Aware personal Information Discovery model based on the cloud. In *2015 Latin American Network Operations and Management Symposium (LANOMS)*. IEEE.

[7] Maria Luisa Damiani. 2014. Location privacy models in mobile applications: conceptual view and research directions. *GeoInformatica* 18, 4 (2014), 819–842.

[8] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports* 1, 3 (2013). https://doi.org/10.1038/srep01376

[9] Apple Differential Privacy Team. 2017. Learning with privacy at scale. (2017).

[10] Khalid A. Eldrandaly, Mohamed Abdel-Basset, and Laila A. Shawky. 2019. Internet of Spatial Things: A New Reference Model With Insight Analysis. *IEEE Access* 7 (2019), 19653–19669. https://doi.org/10.1109/ACCESS.2019.2897012

[11] Jie Feng, Can Rong, Funing Sun, Diansheng Guo, and Yong Li. 2020. PMF: A privacy-preserving human mobility prediction framework via federated learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–21.

[12] Pin-I Han and Hsiao-Ping Tsai. 2015. SST: Privacy preserving for semantic trajectories. In *2015 16th IEEE International Conference on Mobile Data Management*, Vol. 2. IEEE, 80–85.

[13] Fengmei Jin, Wen Hua, Matteo Francia, Pingfu Chao, Maria Orlowska, and Xiaofang Zhou. 2021. A Survey and Experimental Study on Privacy-Preserving Trajectory Data Publishing. (2021).

[14] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021. Advances and Open Problems in Federated Learning. arXiv:1912.04977 [cs.LG]

[15] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36.

[16] Bo Liu, Wanlei Zhou, Tianqing Zhu, Longxiang Gao, and Yong Xiang. 2018. Location privacy and its applications: A systematic study. *IEEE access* 6 (2018).

[17] Adriano Di Luzio, Aline Carneiro Viana, Konstantinos Chatzikokolakis, Georgi Dikov, Catuscia Palamidessi, and Julinda Stefa. 2019. Catch me if you can: how geo-indistinguishability affects utility in mobility-based geographic datasets. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising, LocalRec@SIGSPATIAL 2019, Chicago, Illinois, USA, November 5, 2019*, Panagiotis Bouros, Tamraparni Dasu, Yaron Kanza, Matthias Renz, and Dimitris Sacharidis (Eds.). ACM.

[18] Anna Monreale, Roberto Trasarti, Dino Pedreschi, Chiara Renso, and Vania Bogorny. 2011. C-safety: a framework for the anonymization of semantic trajectories. *Trans. Data Priv.* 4, 2 (2011), 73–101. http://www.tdp.cat/issues11/abs.a077a11.php

[19] Iulian Sandu Popa, Dai Hai Ton That, Karine Zeitouni, and Cristian Borcea. 2019. Mobile participatory sensing with strong privacy guarantees using secure probes. *GeoInformatica* (2019). https://doi.org/10.1007/s10707-019-00389-4 Publisher Copyright: © 2019, Springer Science+Business Media, LLC, part of Springer Nature. Copyright: Copyright 2019 Elsevier B.V., All rights reserved.

[20] Daniele Quercia, Ilias Leontiadis, Liam McNamara, Cecilia Mascolo, and Jon Crowcroft. 2011. SpotME If You Can: Randomized Responses for Location Obfuscation on Mobile Phones. In *2011 International Conference on Distributed Computing Systems, ICDCS 2011, Minneapolis, Minnesota, USA, June 20-24, 2011*. IEEE Computer Society, 363–372. https://doi.org/10.1109/ICDCS.2011.79

[21] Sina Shaham, Ming Ding, Bo Liu, Shuping Dang, Zihuai Lin, and Jun Li. 2020. Privacy preservation in location-based services: a novel metric and attack model. *IEEE Transactions on Mobile Computing* (2020).

[22] Romana Talat, Mohammad S. Obaidat, Muhammad Muzammal, Ali Hassan Sodhro, Zongwei Luo, and Sandeep Pirbhulal. 2020. A decentralised approach to privacy preserving trajectory mining. *Future Gener. Comput. Syst.* 102 (2020), 382–392. https://doi.org/10.1016/j.future.2019.07.068

[23] Dai Hai Ton That, Iulian Sandu Popa, and Karine Zeitouni. 2015. TRIFL: A Generic Trajectory Index for Flash Storage. *ACM Trans. Spatial Algorithms Syst.* 1, 2 (2015), 6:1–6:44. https://doi.org/10.1145/2786758

[24] Nana Wang and Mohan S Kankanhalli. 2020. Protecting sensitive place visits in privacy-preserving trajectory publishing. *Computers & Security* 97 (2020).

[25] Mengmeng Yang, Lingjuan Lyu, Jun Zhao, Tianqing Zhu, and Kwok-Yan Lam. 2020. Local differential privacy and its applications: A comprehensive survey. *arXiv preprint arXiv:2008.03686* (2020).

[26] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. 2021. A survey on federated learning. *Knowledge-Based Systems* 216 (2021), 106775. https://doi.org/10.1016/j.knosys.2021.106775

[27] Yu Zheng. 2015. Trajectory Data Mining: An Overview. *ACM Transaction on Intelligent Systems and Technology* 6, 3, Article 29 (2015).