RESEARCH CENTRE

Inria Saclay Centre
at Institut Polytechnique de Paris

IN PARTNERSHIP WITH:

CNRS, Institut Polytechnique de Paris

2023
ACTIVITY
REPORT

Project-Team

COMETE

Privacy, Fairness and Robustness in Information Management

IN COLLABORATION WITH: Laboratoire d'informatique de l'école polytechnique (LIX)

DOMAIN

Algorithmics, Programming, Software and Architecture

THEME

Security and Confidentiality

Contents

Project-Team COMETE

Creation of the Project-Team: 2008 January 01

Keywords

Computer sciences and digital sciences

A2.1.1. – Semantics of programming languages

A2.1.5. – Constraint programming

A2.1.6. – Concurrent programming

A2.1.9. – Synchronous languages

A2.4.1. – Analysis

A3.4. – Machine learning and statistics

A3.5. – Social networks

A4.1. – Threat analysis

A4.5. – Formal methods for security

A4.8. – Privacy-enhancing technologies

A8.6. – Information theory

A8.11. – Game Theory

A9.1. – Knowledge

A9.2. – Machine learning

A9.7. – AI algorithmics

A9.9. – Distributed AI, Multi-agent

Other research topics and application domains

B6.1. – Software industry

B6.6. – Embedded systems

B9.5.1. – Computer science

B9.6.10. – Digital humanities

B9.9. – Ethics

B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientists

- Catuscia Palamidessi [Team leader, INRIA, Senior Researcher, HDR]
- Frank Valencia [CNRS, Researcher]
- Sami Zhioua [INRIA, Advanced Research Position]

Post-Doctoral Fellows

- Selene Leya Cerna Nahuis [INRIA, Post-Doctoral Fellow, from Feb 2023 until Aug 2023]
- Heber Hwang Arcolezi [INRIA, Post-Doctoral Fellow, until Sep 2023]
- Szilvia Lestyan [INRIA, Post-Doctoral Fellow, until Nov 2023]

PhD Students

- Andreas Athanasiou [INRIA]
- Ruta Binkyte-Sadauskiene [INRIA]
- Sayan Biswas [ECOLE POLY PALAISEAU, from Sep 2023 until Oct 2023]
- Sayan Biswas [INRIA, until Aug 2023]
- Ganesh Del Grosso Guzman [INRIA, until Aug 2023]
- Ramon Goncalves Gonze [INRIA, from Mar 2023]
- Federica Granese [INRIA, until Mar 2023]
- Karima Makhlouf [INRIA]
- Carlos Pinzon Henao [INRIA]

Technical Staff

- Gangsoo Zeong [INRIA, Engineer]

Interns and Apprentices

- Jay Suhas Jawale [INRIA, Intern, from Jun 2023 until Jul 2023]
- Oussama Khammassi [INRIA, Intern, from Jun 2023 until Aug 2023]
- Raluca Panainte [INRIA, Intern, from Jun 2023 until Aug 2023]
- Tamara Stefanovic [INRIA, Intern, from Apr 2023 until Jun 2023]
- Yassine Turki [FX CONSEIL, Intern, from Jun 2023 until Jul 2023]
- Yassine Turki [LIX, Intern, from Feb 2023 until May 2023]

Administrative Assistant

- Mariana De Almeida [INRIA, from Mar 2023]

Visiting Scientists

- Martina Cinquini [UNIV PISE, from Apr 2023 until Jul 2023]

- Josée Desharnais [Laval university, from Sep 2023 until Nov 2023, HDR]

- Daniele Gorla [Università di Roma "La Sapienza", from Mar 2023 until Mar 2023, HDR]

- Annabelle-Kate McIver [UNIV MACQUARIE, from Dec 2023, HDR]

- Charles-Carroll Morgan [UNSW, from Dec 2023, HDR]

External Collaborators

- Konstantinos Chatzikokolakis [CNRS, HDR]

- Mario Sergio Ferreira Alvim Junior [UFMG, HDR]

- Szilvia Lestyan [INED, from Dec 2023]

- Pablo Piantanida [Centrale Supélec, in leave to MILA, Canada, HDR]

2 Overall objectives

The leading objective of COMETE is to develop a principled approach to privacy protection to guide the design of sanitization mechanisms in realistic scenarios. We aim to provide solid mathematical foundations were we can formally analyze the properties of the proposed mechanisms, considered as leading evaluation criteria to be complemented with experimental validation. In particular, we focus on privacy models that:

- allow the sanitization to be *applied and controlled directly by the user*, thus avoiding the need of a trusted party as well as the risk of security breaches on the collected data,
- are *robust with respect to combined attacks*, and
- provide an *optimal trade-off between privacy and utility.*

Two major lines of research are related to machine learning and social networks. These are prominent presences in nowadays social and economical fabric, and constitute a major source of potential problems. In this context, we explore topics related to the propagation of information, like *group polarization*, and other issues arising from the deep learning area, like *fairness* and *robustness with respect to adversarial inputs*, that have also a critical relation with privacy.

3 Research program

The objective of COMETE is to develop principled approaches to some of the concerns in today's technological and interconnected society: privacy, machine-learning-related security and fairness issues, and propagation of information in social networks.

3.1 Privacy

The research on privacy will be articulated in several lines of research.

3.1.1 Three way optimization between privacy and utility

One of the main problems in the design of privacy mechanisms is the preservation of the utility. In the case of local privacy, namely when the data are sanitized by the user before they are collected, the notion of utility is twofold:

Utility as quality of service (QoS): The user usually gives his data in exchange of some service, and in general the quality of the service depends on the precision of such data. For instance, consider a scenario in which Alice wants to use a LBS (Location-Based Service) to find some restaurant near her location x. The LBS needs of course to know Alice's location, at least approximately, in order to provide the service. If Alice is worried about her privacy, she may send to the LBS an approximate location y instead of x. Clearly, the LBS will send a list of restaurants near x, so if y is too far from x the service will degrade, while if it is too close Alice's privacy would be at stake.

Utility as statistical quality of the data (Stat): Bob, the service provider, is motivated to offer his service because in this way he can collect Alice's data, and quality data are very valuable for the big-data industry. We will consider in particular the use of the data collections for statistical purposes, namely for extracting general information about the population (and not about Alice as an individual). Of course, the more Alice's data are obfuscated, the less statistical value they have.

We intend to consider both kinds of utility, and study the "three way" optimization problem in the context of d-privacy, our approach to local differential privacy [45]. Namely, we want to develop methods for producing mechanisms that offer the best trade-off between d-privacy, QoS and Stat, at the same time. In order to achieve this goal, we will need to investigate various issues. In particular:

- how to best estimate the original distribution from a collection of noisy data, in order to perform the intended statistical analysis,
- what metrics to use for assessing the statistical value of a distributions (for a given application), in order to reason about Stat, and
- how to compute in an efficient way the best noise from the point of view of the trade-off between d-privacy, QoS and Stat.

Estimation of the original distribution The only methods for the estimation of the original distribution from perturbed data that have been proposed so far in the literature are the iterative Bayesian update (IBU) and the matrix inversion (INV). The IBU is more general and based on solid statistical principles, but it is not ye well known in the in the privacy community, and it has not been studied much in this context. We are motivated to investigate this method because from preliminary experiments it seems more efficient on date obfuscated by geo-indistinguishability mechanisms (cfr. next section). Furthermore, we believe that the IBU is compositional, namely it can deal naturally and efficiently with the combination of data generated by different noisy functions, which is important since in the local model of privacy every user can, in principle, use a different mechanisms or a different level of noise. We intend to establish the foundations of the IBU in the context of privacy, and study its properties like the compositionality mentioned above, and investigate its performance in the state-of-the-art locally differentially private mechanisms.

Hybrid model An interesting line of research will be to consider an intermediate model between the local and the central models of differential privacy (cfr. Figure 1). The idea is to define a privacy mechanism based on perturbing the data locally, and then collecting them into a dataset organized as an histogram. We call this model "hibrid" because the collector is trusted like in central differential privacy, but the data are sanitized according to the local model. The resulting dataset would satisfy differential privacy from the point of view of an external observer, while the statistical utility would be as high as in the local model. One further advantage is that the IBU is compositional, hence the datasets sanitized in this way could be combined without any loss

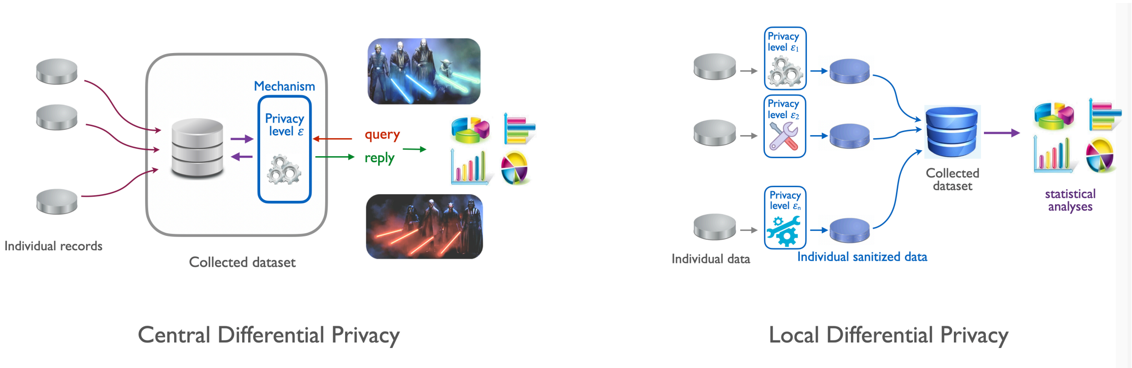
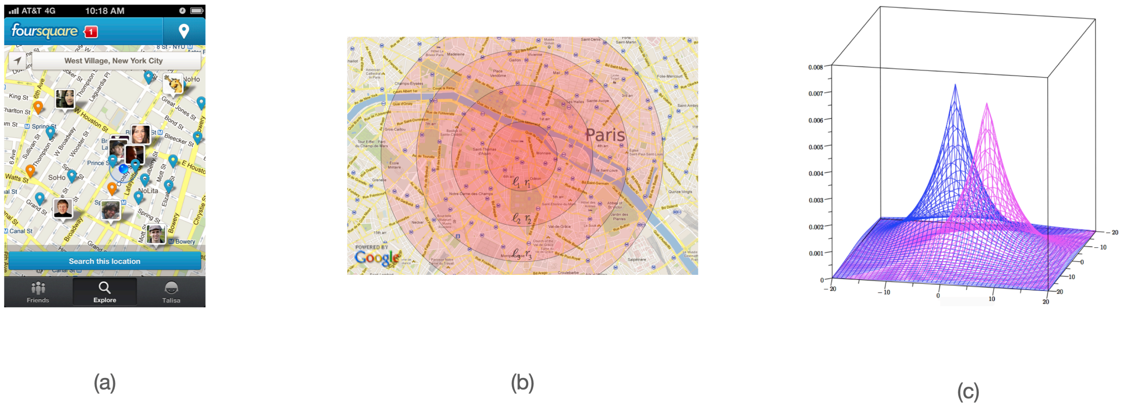

Mechanism
Privacy level ε
query
reply
Individual records
Collected dataset
Central Differential Privacy
Privacy level ε_1
Privacy level ε_2
Privacy level ε_n
Collected dataset
statistical analyses
Individual data
Individual sanitized data
Local Differential Privacy

Figure 1: The central and the local models of differential privacy



(a)

(b)

(c)

Figure 2: Geo-indistinguishability is a framework to protect the privacy of the user when dealing with location-based services (a). The framework guarrantees *d*-privacy, a distance-based variant of differential privacy (b). The typical implementation uses (extended) Laplace noise (c).

of precision in the application of the IBU. In other words, the statistical utility of the union of sanitized datasets is the same as the statistical utility of the sanitized union of datasets, which is of course an improvement (for the law of large numbers) wrt each separate dataset. One important application would be the cooperative sharing of sanitized data owned by different different companies or institution, to the purpose of improving statistical utility while preserving the privacy of their respective datasets.

3.1.2 Geo-indistinguishability

We plan to further develop our line of research on location privacy, and in particular, enhance our framework of geo-indistinguishability [2] (cfr. Figure 2) with mechanisms that allow to take into account sanitize high-dimensional traces without destroying utility (or privacy). One problem with the geo-indistinguishable mechanisms developed so far (the planar Laplace an the planar geometric) is that they add the same noise function uniformly on the map. This is sometimes undesirable: for instance, a user located in a small island in the middle of a lake should generate much more noise to conceal his location, so to report also other locations on the ground, because the adversary knows that it is unlikely that the user is in the water. Furthermore, for the same reason, it does not offer a good protection with respect to re-identification attacks: a user who lives in an isolated place, for instance, can be easily singled out because he reports locations far away from all others. Finally, and this is a common problem with all methods based on DP, the repeated use of the mechanism degrades the privacy, and even when the degradation is linear, as in the case of all DP-based methods, it becomes quickly unacceptable when dealing with highly structured data such

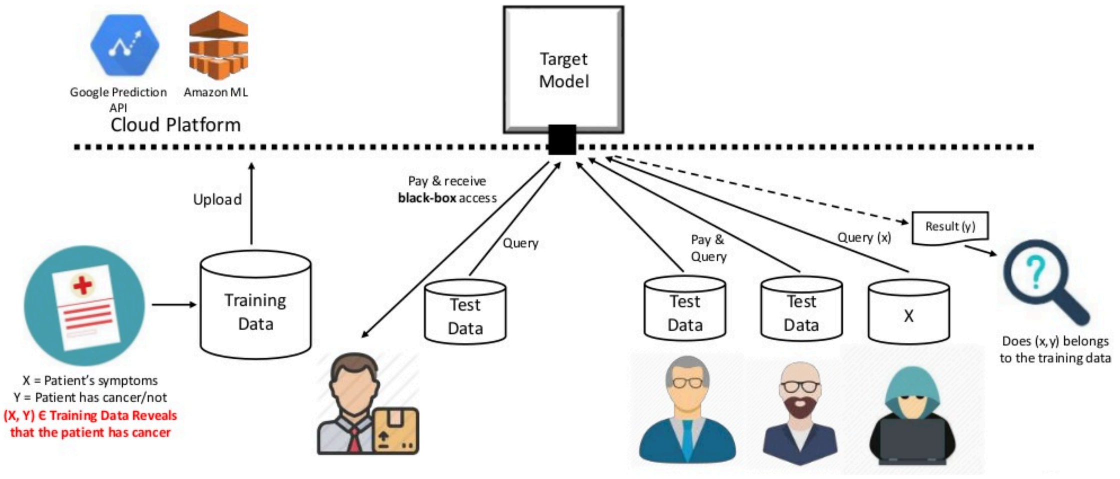


Figure 3: Privacy breach in machine learning as a service.

as spatio-temporal traces.

3.1.3 Threats for privacy in machine learning

In recent years several researchers have observed that machine learning models leak information about the training data. In particular, in certain cases an attacker can infer with relatively high probability whether a certain individual participated in the dataset (*membership inference attack*) od the value of his data (*model inversion attack*). This can happen even if the attacker has nop access to the internals of the model, i.e., under the *black box assumption*, which is the typical scenario when machine learning is used as a service (cfr. Figure 3). We plan to develop methods to reason about the information-leakage of training data from deep learning systems, by identifying appropriate measures of leakage and their properties, and use this theoretical framework as a basis for the analysis of attacks and for the development of robust mitigation techniques. More specifically, we aim at:

- Developing compelling case studies based on state-of-the-art algorithms to perform attacks, showcasing the feasibility of uncovering specified sensitive information from a trained software (model) on real data.

- Quantifying information leakage. Based on the uncovered attacks, the amount of sensitive information present in trained software will be quantified and measured. We will study suitable notions of leakage, possibly based on information-theoretical concepts, and establish firm foundations for these.

- Mitigating information leakage. Strategies will be explored to avoid the uncovered attacks and minimize the potential information leakage of a trained model.

3.1.4 Relation between privacy and robustness in machine learning

The relation between privacy and robustness, namely resilience to adversarial attacks, is rather complicated. Indeed the literature on the topic seems contradictory: on the one hand, there are works that show that differential privacy can help to mitigate both the risk of inference attacks and of misclassification (cfr. [52]). On the other hand, there are studies that show that there is a trade-off between protection from inference attacks and robustness [54]. We intend to shed light on this confusing situation. We believe that the different variations of differential privacy play a role in this apparent contradiction. In particular, *preprocessing* the training data with *d*-privacy seems to go along with the concept of robustness, because it guarantees that small variations in the input cannot result in large variations in the output, which is exactly the principle of robustness. On the other hand, the addition of random noise on the output result (*postprocessing*), which is the typical method in central DP, should reduce the precision and therefore increase the possibility of misclassification. We intend to make a taxonomy of the differential privacy variants, in relation to their effect on robustness, and develop a principled approach to protect both privacy and security in an optimal way.

One promising research direction for the deployment of d-privacy in this context is to consider Bayesian neural networks (BNNs). These are neural networks with distributions over their weights, which can capture the uncertainty within the learning model, and which provide a natural notion of distance (between distributions) on which we can define a meaningful notion of d-privacy. Such neural networks allow to compute an uncertainty estimate along with the output, which is important for safety-critical applications.

3.1.5 Relation between privacy and fairness

Both fairness and privacy are multi-faces notions, assuming different meaning depending on the application domain, on the situation, and on what exactly we want to protect. Fairness, in particular, has received many different definitions, some even in contrast with each other. One of the definitions of fairness is the property that similar “similar” input data produce "similar" outputs. Such notion corresponds closely to d-privacy. Other notions of fairness, however, are in opposition to standard differential privacy. This is the case, notably, of *Equalized Odds* [47] and of *Equality of False Positives* and *Equality of False Negatives* [46]. We intend to study a tassonomy of the relation between the main notions of fairness an the various variants of differential privacy. In particular, we intend to study the relation between the recently-introduced notions of *causal fairness* and *causal differential privacy* [55].

Another line of research related to privacy and fairness, that we intend to explore, is the design of to pre-process the training set so to obtain machine learning models that are both privacy-friendly and fair.

3.2 Quantitative information flow

In the area of quantitive information flow (QIF), we intend to pursue two lines of research: the study of non-0-sum games, and the estimation of g-leakage [43] under the black-box assumption.

3.2.1 Non-0-sum games

The framework of g-leakage does not take into account two important factors: (a) the loss of the user, and (b) the cost of the attack for the adversary. Regarding (a), we observe that in general the goal of the adversary may not necessarily coincide with causing maximal damage to the user, i.e., there may be a mismatch between the aims of the attacker and what the user tries to protect the most. To model this more general scenario, we had started investigating the interplay between defender and attacker in a game-theoretic setting, starting with the simple case of 0-sum games which corresponds to g-leakage. The idea was that, once the simple 0-sum case would be well understood, we would extend the study to the non-0-sum case, that is needed to represent (a) and (b) above. However, we had first to invent and lay the foundations of a new kind of games, the *information leakage games* [42] because the notion of leakage cannot be expressed in terms of payoff in standard game theory. Now that the theory of these new games is well established, we intend to go ahead with our plan, namely study costs and damages of attacks in terms of non-0-sum information leakage games.

3.2.2 Black-box estimation of leakage via machine learning

Most of the works in QIF rely on the so-called white-box assumption, namely, they assume that it is possible to compute exactly the (probabilistic) input-output relation of the system, seen as an information-theoretic channel. This is necessary in order to apply the formula that expresses the leakage. In practical situations, however, it may not be possible to compute the input-output relation, either because the system is too complicated, or simply because it is not accessible. Such scenario is called black-box. The only assumption we make is that the adversary can interact with the system, by feeding to it inputs of his choice and observing the corresponding outputs.

Given the practical interest of the black-box model, we intend to study methods to estimate its leakage. Clearly the standard QIF methods are not applicable. We plan to use, instead, a machine learning approach, continuing the work we started in [10]. In particular, we plan to investigate whether we can improve the efficiency of the method proposed by leveraging on the experience that we have acquired with the GANs [53]. The idea is to construct a training set and a testing set from the input-output samples collected by interacting with the system, and then build a classifier that learns from the training set to classify the input from the output so to maximize its gain. The measure of its performance on the testing set should then give an estimation of the posterior g-vulnerability.

3.3 Information leakage, bias and polarization in social networks

One of the core activities of the team will be the study of how information propagate in the highly interconnected scenarios made possible by modern technologies. We will consider the issue of privacy protection as well as the social impact of privacy leaks. Indeed, recent events have shown that social networks are exposed to actors malicious agents that can collect *private information* of millions of users with or without their consent. This information can be used to build psychological profiles for microtargeting, typically aimed at discovering users preconceived beliefs and at reinforcing them. This may result in polarization of opinions as people with opposing views would tend to interpret new information in a biased way causing their views to move further apart. Similarly, a group with uniform views often tends to make more extreme decisions than its individual. As a result, users may become more radical and isolated in their own ideological circle causing dangerous splits in society.

3.3.1 Privacy protection

In [50] we have investigated potential leakage in social networks, namely, the unintended propagation and collection of confidential information. We intend to enrich this model with epistemic aspects, in order to take into account the belief of the users and how it influences the behavior of agents with respect the transmission of information.

Furthermore, we plan to investigate attack models used to reveal a user's private information, and explore the framework of g-leakage to formalize the privacy threats. This will provide the basis to study suitable protection mechanisms.

3.3.2 Polarization and Belief in influence graphs

In social scenarios, a group may shape their beliefs by attributing more value to the opinions of influential figures. This cognitive bias is known as *authority bias*. Furthermore, in a group with uniform views, users may become extreme by reinforcing one another's opinions, giving more value to opinions that confirm their own beliefs; another common cognitive bias known as *confirmation bias*. As a result, social networks can cause their users to become radical and isolated in their own ideological circle causing dangerous splits in society (polarization). We intend to study these dynamics in a model called *influence graph*, which is a weighted directed graph describing connectivity and influence of each agent over the others. We will consider two kinds of belief updates: the authority belief update, which gives more value to the opinion of agents with higher influence, and the confirmation bias update, which gives more value to the opinion of agents with similar views.

We plan to study the evolution of polarization in these graphs. In particular, we aim at defining a suitable measure of polarization, characterizing graph structures and conditions under which polarization eventually converges to 0 (vanishes), and methods to compute the change in the polarization value over time.

Another purpose of this line of research is how the bias of the agents whose data are being collected impacts the *fairness* of learning algorithms based on these data.

3.3.3 Concurrency models for the propagation of information

Due to their popularity and computational nature, social networks have exacerbated group polarization. Existing models of group polarization from economics and social psychology state its basic principles and measures [48]. Nevertheless, unlike our computational ccp models, they are not suitable for describing the dynamics of agents in distributed systems. Our challenge is to coherently combine our ccp models for epistemic behavior with principles and techniques from economics and social psychology for GP. We plan to develop a ccp-based process calculus which incorporates structures from social networks, such as communication, influence, individual opinions and beliefs, and privacy policies. The expected outcome is a *computational model* that will allow us to specify the interaction of groups of agents exchanging *epistemic information* among them and to predict and measure the *leakage of private information*, as well as the *degree of polarization* that such group may reach.

4 Application domains

The application domains of our research include the following:

Protection of sensitive personal data Our lives are growingly entangled with internet-based technologies and the limitless digital services they provide access to. The ways we communicate, work, shop, travel, or entertain ourselves are increasingly depending on these services. In turn, most such services heavily rely on the collection and analysis of our personal data, which are often generated and provided by ourselves: tweeting about an event, searching for friends around our location, shopping online, or using a car navigation system, are all examples of situations in which we produce and expose data about ourselves. Service providers can then gather substantial amounts of such data at unprecedented speed and at low cost.

While data-driven technologies provide undeniable benefits to individuals and society, the collection and manipulation of personal data has reached a point where it raises alarming privacy issues. Not only the experts, but also the population at large are becoming increasingly aware of the risks, due to the repeated cases of violations and leaks that keep hitting the headlines. Examples abound, from iPhones storing and uploading device location data to Apple without users' knowledge to the popular Angry Birds mobile game being exploited by NSA and GCHQ to gather users' private information such as age, gender and location.

If privacy risks connected to personal data collection and analysis are not addressed in a fully convincing way, users may eventually grow distrustful and refuse to provide their data. On the other hand, misguided regulations on privacy protection may impose excessive restrictions that are neither necessary nor sufficient. In both cases, the risk is to hinder the development of many high-societal-impact services, and dramatically affect the competitiveness of the European industry, in the context of a global economy which is more and more relying on Big Data technologies.

The EU General Data Protection Regulation (GDPR) imposes that strong measures are adopted by-design and by-default to guarantee privacy in the collection, storage, circulation and analysis of personal data. However, while regulations set the high-level goals in terms of privacy, it remains an open research challenge to map such high-level goals into concrete requirements and to develop privacy-preserving solutions that satisfy the legally-driven requirements. The current de-facto standard in personal data sanitization used in the industry is anonymization (i.e., personal identifier removal or substitution by a pseudonym). Anonymity however does not offer any actual protection because of potential *linking attacks* (which have actually been known since a long time). Recital 26 of the GDPR states indeed that anonymization may be insufficient and that anonymized data must still be treated as personal data. However the regulation provide no guidance on how or what constitutes an effective data re-identification scheme, leaving a grey area on what could be considered as adequate sanitization.

In COMETE, we pursue the vision of a world where pervasive, data-driven services are inalienable life enhancers, and at the same time individuals are fully guaranteed that the privacy of their sensitive personal data is protected. Our objective is to develop a principled approach to the design of sanitization mechanisms providing an optimal trade-off between privacy and utility, and robust with respect to composition attacks. We aim at establishing solid mathematical foundations were we can formally analyze the properties of the proposed mechanisms, which will be regarded as leading evaluation criteria, to be complemented with experimental validation.

We focus on privacy models where the sanitization can be applied and controlled directly by the user, thus avoiding the need of a trusted party as well as the risk of security breaches on the collected data.

Ethical machine learning Machine learning algorithms have more and more impact on and in our day-to-day lives. They are already used to take decisions in many social and economical domains, such as recruitment, bail resolutions, mortgage approvals, and insurance premiums, among many others. Unfortunately, there are many ethical challenges:

- Lack of transparency of machine learning models: decisions taken by these machines are not always intelligible to humans, especially in the case of neural networks.

- Machine learning models are not neutral: their decisions are susceptible to inaccuracies, discriminatory outcomes, embedded or inserted bias.

- Machine learning models are subject to privacy and security attacks, such as data poisoning and membership and attribiute inference attacks.

The time has therefore arrived that the most important area in machine learning is the implementation of algorithms that adhere to ethical and legal requirements. For example, the United States' Fair Credit Reporting Act and European Union's General Data Protection Regulation (GDPR) prescribe that data must be processed in a way that is fair/unbiased. GDPR also alludes to the right of an individual to receive an explanation about decisions made by an automated system.

One of the goals of COMETE's research is to contribute to make the machine learning technology evolve towards compliance with the human principles and rights, such as fairness and privacy, while continuing to improve accuracy and robustness.

Polarization in Social Networks *Distributed systems* have changed substantially with the advent of social networks. In the previous incarnation of distributed computing the emphasis was on consistency, fault tolerance, resource management and other related topics. What marks the new era of distributed systems is an emphasis on the flow of *epistemic* information (knowledge, facts, opinions,beliefs and lies) and its impact on democracy and on society at large.

Indeed in social networks a group may shape their beliefs by attributing more value to the opinions of influential figures. This cognitive bias is known as *authority bias*. Furthermore, in a group with uniform views, users may become extreme by reinforcing one another's opinions, giving more value to opinions that confirm their own beliefs; another common cognitive bias known as *confirmation bias*. As a result, social networks can cause their users to become radical and isolated in their own ideological circle causing dangerous splits in society in a phenomenon known as *polarization*.

One of our goals in COMETE is to study the flow of epistemic information in social networks and its impact on opinion shaping and social polarization. We study models for reasoning about distributed systems whose agents interact with each other like in social networks; by exchanging epistemic information and interpreting it under different biases and network topologies. We are interested in predicting and measuring the degree of polarization that such agents may reach. We focus on polarization with strong influence in politics such as affective polarization; the dislike and distrust those from the other political party. We expect the model to provide social networks with guidance as to how to distribute newsfeed to mitigate polarization.

5 Social and environmental responsibility

5.1 Footprint of research activities

Whenever possible, the members of COMETE have privileged attendance of conferences and workshops on line, to reduce the environmental impact of traveling.

6 Highlights of the year

6.1 Awards

- Test of Time award at the 2023 ACM Conference on Computer and Communications Security (CCS), for the paper [2] published in CCS'13.
- Catuscia palamidessi has received a Honorable Mention for the Bozenna Pasik-Duncan Mentorship Award at the IEEE Returning Mothers conference, 2023. This prize is meant to reward mentors who help people to return to their studies. Indeed, our team has welcome as PhD students some women who were married and with children (Natasha Fernandes, Ruta Bynkite and Karima Makhlouf), and has helped some young men who had crises during their PhD to get back in track.
- Best paper award at DBSec 2023 for the paper [23].
- Best paper award at CADE 2023 for the paper [21].
- Carlos Pinzon Henao received the accessit to the 2023 Doctoral Prize awarded by the University of Paris Saclay and the Polytechnic Institute of Paris, for his work [9].

6.2 New funded projects

- The ANR project DIFPRIPOS, for which Catuscia Palamidessi is the PI for Inria Saclay, has been accepted.

6.3 Organization of events

- Catuscia Palamidessi has co-organized and co-chaired the Fourth AAAI Workshop on Privacy-Preserving Artificial Intelligence. Washington DC, USA. February 2023.
- Szilvia Lestyan, Ruta Binkyte-Sadauskiene, Ramon Goncalves Gonze, Catuscia Palamidessi, and Bibel Benbouzid have organized and chaired the Second Ethical AI @Comete workshop. Palaiseau, France. November 2023.
- Frank Valencia has organized and chaired the Promueva Workshop on Models for Social Networks at École Polytechnique, Sorbonne Paris Nord, and IRCAM. June 12-23, 2023.
- Frank Valencia has co-organized and co-chaired the Promueva Public Workshop on Polarization in Social Networks and AI Impact, at Universidad Javeriana Cali, August 23, 2023.
- Héber Hwang Arcolezi has co-organized the 13th French Workshop on Privacy, at University Bourgogne Franche-Comté, June 12-15, 2023.

7 New software, platforms, open data

7.1 New software

7.1.1 Multi-Freq-LDPy

Name: Multiple Frequency Estimation Under Local Differential Privacy in Python

Keywords: Privacy, Python, Benchmarking

Scientific Description: The purpose of Multi-Freq-LDPy is to allow the scientific community to benchmark and experiment with Locally Differentially Private (LDP) frequency (or histogram) estimation mechanisms. Indeed, estimating histograms is a fundamental task in data analysis and data mining that requires collecting and processing data in a continuous manner. In addition to the standard single frequency estimation task, Multi-Freq-LDPy features separate and combined multidimensional and longitudinal data collections, i.e., the frequency estimation of multiple attributes, of a single attribute throughout time, and of multiple attributes throughout time.

Functional Description: Local Differential Privacy (LDP) is a gold standard for achieving local privacy with several real-world implementations by big tech companies such as Google, Apple, and Microsoft. The primary application of LDP is frequency (or histogram) estimation, in which the aggregator estimates the number of times each value has been reported.

Multi-Freq-LDPy provides an easy-to-use and fast implementation of state-of-the-art LDP mechanisms for frequency estimation of: single attribute (i.e., the building blocks), multiple attributes (i.e., multidimensional data), multiple collections (i.e., longitudinal data), and both multiple attributes/collections.

Multi-Freq-LDPy is now a stable package, which is built on the well-established Numpy package - a de facto standard for scientific computing in Python - and the Numba package for fast execution.

URL: https://github.com/hharcolezi/multi-freq-ldpy

Publication: hal-03816212

Contact: Heber Hwang Arcolezi

Participants: Heber Hwang Arcolezi, Jean-François Couchot, Sebastien Gambs, Catuscia Palamidessi, Majid Zolfaghari

7.1.2 LOLOHA

Name: LOngitudinal LOcal HAshing For Locally Private Frequency Monitoring

Keyword: Privacy

Functional Description: This is a Python implementation of our locally differentially private mechanism named LOLOHA. We implemented a private-oriented version named BiLOLOHA and a utility-oriented version named OLOLOHA. We benchmarked our mechanisms in comparison with Google's RAPPOR mechanism and Microsoft's dBitFlipPM mechanism.

URL: https://github.com/hharcolezi/LOLOHA

Publication: hal-03911550

Contact: Heber Hwang Arcolezi

Participants: Heber Hwang Arcolezi, Sebastien Gambs, Catuscia Palamidessi, Carlos Pinzon Henao

7.1.3 PRiLDP

Name: Privacy Risks of Local Differential Privacy

Keyword: Privacy

Functional Description: This is a Python implementation of two privacy threats we identified against locally differentially private (LDP) mechanisms. We implemented attribute inference attacks as well as re-identification attacks, benchmarking the robustness of five state-of-the-art LDP mechanisms.

URL: `https://github.com/hharcolezi/risks-ldp`

Publication: hal-04082592

Contact: Heber Hwang Arcolezi

Participants: Heber Hwang Arcolezi, Sebastien Gambs, Jean-François Couchot, Catuscia Palamidessi

7.1.4 PRIVIC

Name: A privacy-preserving method for incremental collection of location data

Keyword: Privacy

Functional Description: This library contains various tools for the PRIVIC project: the implementation of the Blahut-Arimoto mechanism for metric privacy, the Iterative Bayesian Update, and the implementation of an algorithm performing an incremental collection of data under metric differential privacy protection, and gradual improvement of the mechanism from the point of view of utility.

URL: `https://github.com/blitzwas/PRIVIC`

Publication: hal-03968692

Contact: Sayan Biswas

Participants: Sayan Biswas, Catuscia Palamidessi

7.1.5 LDP-FAIRNESS

Name: Impact of Local Differential Privacy on Fairness

Keywords: Privacy, Fairness

Functional Description: This library contains various tools for the study of the impact of Local Differential Privacy on fairness.

URL: `https://github.com/hharcolezi/ldp-fairness-impact`

Publication: hal-04175027

Contact: Heber Hwang Arcolezi

Participants: Heber Hwang Arcolezi, Karima Makhlouf, Catuscia Palamidessi

7.1.6 Causal-based Fairness

Name: Causal-based Machine Learning Discrimination Estimation

Keywords: Fairness, Causal discovery

Functional Description: Addressing the problem of fairness is crucial to safely use machine learning algorithms to support decisions with a critical impact on people's lives such as job hiring, child maltreatment, disease diagnosis, loan granting, etc. Several notions of fairness have been defined and examined in the past decade, such as statistical parity and equalized odds. The most recent fairness notions, however, are causal-based and reflect the now widely accepted idea that using causality is necessary to appropriately address the problem of fairness. The big impediment to the use of causality to address fairness, however, is the unavailability of the causal model (typically represented as a causal graph). This library contains the software tools that implement all required steps to estimate discrimination using a causal approach, including, the causal discovery, the adjustment of the causal model, and the estimation of discrimination. The software is to be deployed as a web application which makes it accessible online without any required setup on the user side.

Publication: hal-04355882

Contact: Sami Zhioua

Participants: Raluca Panainte, Yassine Turki, Sami Zhioua

7.1.7 Polarization

Name: A model for polarization

Keyword: Social network

Functional Description: This is a Python implementation of our polarization model. The implementation is parametric in the social influence graph and belief update representing the social network and it allows for the simulation of belief evolution and measuring the polarization of the network.

URL: `https://github.com/Sirquini/Polarization`

Publication: hal-03872692

Contact: Frank Valencia

Participants: Frank Valencia, Mario Ferreira Alvim Junior, Sophia Knight, Santiago Quintero

7.1.8 GMeet

Name: GMeet Algorithms

Keyword: Distributed computing

Functional Description: This is a Python library containing the implementation of our methods to compute distributed knowledge in multi-agent systems. The implementation allows for experimental comparison between the different methods on randomly generated inputs.

URL: `https://caph1993.github.io/GMeetMono/`

Publication: hal-02422624

Contact: Frank Valencia

7.1.9 Fairness-Accuracy

Name: On the trade-off between Fairness and Accuracy

Keywords: Fairness, Machine learning

Functional Description: This software is composed by two main modules that serve the following purposes:

(1) To visualize the perimeter of all possible machine learning models in the Equal Opportunity - Accuracy space, and to show that, for certain distributions, Equal Opportunity implies that the best Accuracy achievable is that of a trivial model.

(2) To compute the Pareto optimality between Equal Opportunity Difference and Accuracy.

Publication: hal-04308195

Contact: Catuscia Palamidessi

Participants: Carlos Pinzon Henao, Catuscia Palamidessi, Pablo Piantanida, Frank Valencia

8 New results

Participants: Catuscia Palamidessi, Frank Valencia, Sami Zhioua, Heber Hwang Arcolezi, Gangsoo Zeong, Sayan Biswas, Ruta Binkyte-Sadauskiene, Ganesh Del Grosso, Federica Granese, Karima Makhlouf, Carlos Pinzon Henao.

8.1 Metric Differential Privacy for Location Data

Location data have been shown to carry a substantial amount of sensitive information. A standard method to mitigate the privacy risks for location data consists in adding noise to the true values to achieve geo-indistinguishability (geo-ind), [2]. However, geo-ind alone is not sufficient to cover all privacy concerns. In particular, isolated locations are not sufficiently protected by the state-of-the-art Laplace mechanism (LAP) for geo-ind. In [14], we have proposed a mechanism based on the Blahut-Arimoto algorithm (BA) from the rate-distortion theory. We have showed that BA, in addition to providing geo-ind, enforces an elastic metric that mitigates the problem of isolation. Furthermore, BA provides an optimal trade-off between information leakage and quality of service. We have also studied the utility of BA in terms of the statistics that can be derived from the reported data, focusing on the inference of the original distribution. To this purpose, we de-noise the reported data by applying the iterative Bayesian update (IBU), an instance of the expectation-maximization method. It turns out that BA and IBU are dual to each other, and as a result, they work well together, in the sense that the statistical utility of BA is quite good and better than LAP for high privacy levels. Exploiting these properties of BA and IBU, we have proposed an iterative method, PRIVIC, for a privacy-friendly incremental collection of location data from users by service providers. We have illustrated the soundness and functionality of our method both analytically and with experiments.

8.2 Robust utility bounds for metric differential privacy

Ghosh et al. introduced in [49] the idea of universal optimality to characterise the “best” mechanism for a certain query that simultaneously satisfies (a fixed) ϵ-differential privacy constraint whilst at the same time providing better utility compared to any other ϵ-differentially private mechanism for the same query. They showed that the Geometric mechanism is universally optimal for the class of counting queries. On the other hand, Brenner and Nissim showed in [44] that outside the space of counting queries, and for the Bayes risk loss function, no such universally optimal mechanisms exist. Except for the universal optimality of the Laplace mechanism, there have been no generalisations of these universally optimal results to other classes of differentially-private mechanisms. In [17], we have used metric differential privacy and quantitative information flow as the fundamental principle for studying universal optimality. Metric differential privacy is a generalisation of both standard (i.e., central) differential privacy and local differential privacy, and it is increasingly being used in various application domains, for instance in location privacy and in privacy-preserving machine learning. Similar to the approaches adopted by Ghosh et al. and Brenner and Nissim, we have measured utility in terms of loss functions, and we have interpreted the notion of a privacy mechanism as an information-theoretic channel satisfying constraints defined by ϵ-differential privacy and a metric meaningful to the underlying state space. Using this framework we were able to clarify Nissim and Brenner's negative results by (a) that in fact all privacy types contain optimal mechanisms relative to certain kinds of non-trivial loss functions, and (b) extending and generalising their negative results beyond Bayes risk specifically to a wide class of non-trivial loss functions. Our exploration suggests that universally optimal mechanisms are indeed rare within privacy types. We therefore propose weaker universal benchmarks of utility called privacy type capacities. We have shown that such capacities always exist and can be computed using a convex optimisation algorithm. Further, we have illustrated these ideas on a selection of examples with several different underlying metrics.

8.3 Utility gain of iterative Bayesian update for Locally differentially private mechanisms

In [22] we have investigated the utility gain of using Iterative Bayesian Update (IBU) for private discrete distribution estimation using data obfuscated with Locally Differentially Private (LDP) mechanisms. We have compared the performance of IBU to Matrix Inversion (MI), a standard estimation technique, for seven LDP mechanisms designed for onetime data collection and for other seven LDP mechanisms designed for multiple data collections (e.g., RAPPOR). To broaden the scope of our study, we have also varied the utility metric, the number of users n, the domain size k, and the privacy parameter ϵ, using both synthetic and real-world data. Our results suggest that IBU can be a useful post-processing tool for improving the utility of LDP mechanisms in different scenarios without any additional privacy cost. For instance, our experiments show that IBU can provide better utility than MI, especially in high privacy regimes (i.e., when ϵ is small). Our paper provides insights for practitioners to use IBU in conjunction with existing LDP mechanisms for more accurate and privacy-preserving data analysis. Finally, we have implemented IBU for all fourteen LDP mechanisms into the state-of-the-art multi-freq-ldpy Python package (link) and open-sourced all our code used for the experiments as tutorials.

8.4 The shuffle model

The shuffle model is an intermediate paradigm between the central and the local models of differential privacy (DP), and it has recently gained popularity. As an initial step, the shuffle model uses a local mechanism to perturb the data individually like the local model of DP. After this local sanitization, a shuffler uniformly permutes the noisy data to dissolve their links with the corresponding data providers. This allows the shuffle model to achieve a certain level of DP guarantee using less noise than the local model, thus providing a better utility for the same level of privacy.

However, the privacy implications of shuffling are not always immediately evident, and derivations of privacy bounds are made on a case-by-case basis. In [29], we have analyzed the combination of LDP with shuffling in the rigorous framework of quantitative information flow (QIF), and have studied the resulting resilience to inference attacks. QIF naturally captures (combinations of) randomization mechanisms as information-theoretic channels, thus allowing for precise modeling of a variety of inference attacks in a natural way and for measuring the leakage of private information under these attacks. We have exploited symmetries of the particular combination of k-RR mechanisms with the shuffle model to achieve closed formulas that express leakage exactly. In particular, we have provided formulae that show how shuffling improves protection against leaks in the local model, and have studied how leakage behaves for various values of the privacy parameter of the LDP mechanism. In contrast to the strong adversary from differential privacy, who knows everyone's record in a dataset but the target's, we have focused on an uninformed adversary, who does not know the value of any individual in the dataset. This adversary is often more realistic as a consumer of statistical datasets, and indeed we have showed that in some situations mechanisms that are equivalent w.r.t. the strong adversary can provide different privacy guarantees under the uninformed one. Finally, we have also illustrated the application of our model to the typical strong adversary from DP.

8.5 Bayes security

Security system designers favor worst-case security metrics, such as those derived from differential privacy (DP), due to the strong guarantees they provide. On the downside, these guarantees result in a high penalty on the system's performance. In [26], we have studied Bayes security, a security metric inspired by the cryptographic advantage. Similarly to DP, Bayes security (a) is independent of an adversary's prior knowledge, (b) it captures the worst-case scenario for the two most vulnerable secrets (e.g., data records); and (c) it is easy to compose, facilitating security analyses. Additionally, Bayes security (d) can be consistently estimated in a black-box manner, contrary to DP, which is useful when a formal analysis is not feasible; and (e) provides a better utility-security trade-off in high-security regimes because it quantifies the risk for a specific threat model as opposed to threat-agnostic metrics such as DP. We have formulated a theory around Bayes security, and we have provided a thorough comparison with respect to well-known metrics, identifying the scenarios where Bayes Security is advantageous for designers.

8.6 Obfuscation padding schemes

In [30] we have considered a set of users, each of which is choosing and downloading one file out of a central pool of public files, and an attacker that observes the download size for each user to identify the choice of each user. We have studied the problem of padding the files so to obfuscate the exact file sizes and minimize the expected accuracy of the attacker, without exceeding some given padding constraints. We have derived the algorithm that finds the optimal padding scheme, proved its correctness, and compared it with an existing solution that uses a different attack model. We have also discussed how the two solutions are related in terms of private information leakage.

8.7 Collecting multidimensional data under local differential privacy

The private collection of multiple statistics from a population is a fundamental statistical problem. One possible approach to realize this is to rely on the local model of differential privacy (LDP). Numerous LDP protocols have been developed for the task of frequency estimation of single and multiple attributes. These studies mainly focused on improving the utility of the algorithms to ensure the server performs the estimations accurately. In [13], we have investigated privacy threats (re-identification and attribute inference attacks) against LDP protocols for multidimensional data following two state-of-the-art solutions for frequency estimation of multiple attributes. To broaden the scope of our study, we have also experimentally assessed five widely used LDP protocols, namely, generalized randomized response, optimal local hashing, subset selection, RAPPOR and optimal unary encoding. Finally, we have also proposed a countermeasure that improves both utility and robustness against the identified threats. Our contributions can help practitioners aiming to collect users' statistics privately to decide which LDP mechanism best fits their needs.

8.8 Frequency estimation of evolving data under local differential privacy

Collecting and analyzing evolving longitudinal data has become a common practice. One possible approach to protect the users' privacy in this context is to use local differential privacy (LDP) protocols, which ensure the privacy protection of all users even in the case of a breach or data misuse. Existing LDP data collection protocols such as Google's RAPPOR and Microsoft's dBitFlipPM can have longitudinal privacy linear to the domain size k, which is excessive for large domains, such as Internet domains. To solve this issue, in [24] we have introduced a new LDP data collection protocol for longitudinal frequency monitoring named LOngitudinal LOcal HAshing (LOLOHA) with formal privacy guarantees. In addition, the privacy-utility trade-off of our protocol is only linear with respect to a reduced domain size g. LOLOHA combines a domain reduction approach via local hashing with double randomization to minimize the privacy leakage incurred by data updates. As demonstrated by our theoretical analysis as well as our experimental evaluation, LOLOHA achieves a utility competitive to current state-of-the-art protocols, while substantially minimizing the longitudinal privacy budget consumption by up to k/g orders of magnitude.

8.9 Bounding information leakage in machine learning

It is well known that Machine Learning models can leak sensitive information about their training data. This information leakage can give rise to membership and attribute inference attacks. Although many attack strategies have been proposed, little effort has been made to formalize these problems. In [15] we have proposed a novel formalism, generalizing membership and attribute inference attack setups previously studied in the literature and connecting them to memorization and generalization. First, we have derived a universal bound on the success rate of inference attacks and connect it to the generalization gap of the target model. Second, we have studied the question of how much sensitive information is stored by the algorithm about its training set and we derive bounds on the mutual information between the sensitive attributes and model parameters. We then have illustrated, experimentally, the potential of our approach by applying it to both synthetic data and classification tasks on natural images. Finally, we have applied our formalism to different attribute inference strategies, with which an adversary is able to recover the identity of writers in the PenDigits dataset.

8.10 Group privacy for personalized federated learning

Federated learning (FL) is a type of collaborative machine learning where participating peers/clients process their data locally, sharing only updates to the collaborative model. This enables to build privacy-aware distributed machine learning models, among others. The goal is the optimization of a statistical model's parameters by minimizing a cost function of a collection of datasets which are stored locally by a set of clients. This process exposes the clients to two issues: leakage of private information and lack of personalization of the model. On the other hand, with the recent advancements in various techniques to analyze data, there is a surge of concern for the privacy violation of the participating clients. To mitigate this, differential privacy and its variants serve as a standard for providing formal privacy guarantees. Often the clients represent very heterogeneous communities and hold data which are very diverse. Therefore, aligned with the recent focus of the FL community to build a framework of personalized models for the users representing their diversity, it is also of utmost importance to protect the clients' sensitive and personal information against potential threats. To address this goal, in [27, 18] we consider metric differential privacy (metric DP) [45], which has the advantage of using a metric-based obfuscation technique that preserves the topological distribution of the original data. To cope with the issue of protecting the privacy of the clients and allowing for personalized model training to enhance the fairness and utility of the system, we have propose a method to provide group privacy guarantees exploiting some key properties of metric DP which enables personalized models under the framework of FL. This method, besides enabling personalized model training in a federated approach and providing formal privacy guarantees, possesses significantly better group fairness measured under a variety of standard metrics than a global model trained within a classical FL template. Theoretical justifications for the applicability are provided, as well as experimental validation on real-world datasets to illustrate the effectiveness of the proposed method.

8.11 Local differential privacy and fairness

In recent years, Local Differential Privacy (LDP), a robust privacy-preserving methodology, has gained widespread adoption in realworld applications. With LDP, users can perturb their data on their devices before sending it out for analysis. However, as the collection of multiple sensitive information becomes more prevalent across various industries, collecting a single sensitive attribute under LDP may not be sufficient. Correlated attributes in the data may still lead to inferences about the sensitive attribute. In [23] we have conducted an empirical study of the impact on fairness of the application of LDP to sensitive attributes. We have proposed a novel privacy budget allocation scheme that considers the varying domain size of sensitive attributes. This generally led to a better privacy-utility-fairness trade-off in our experiments than the state-of-art solution. Our results show that LDP leads to slightly improved fairness in learning problems without significantly affecting the performance of the models. We conduct extensive experiments evaluating three benchmark datasets using several group fairness metrics and seven state-of-the-art LDP protocols. Overall, this study challenges the common belief that differential privacy necessarily leads to worsened fairness in machine learning.

8.12 Fairness and accuracy

One of the main concerns about fairness in machine learning (ML) is that, in order to achieve it, one may have to trade off some accuracy. To overcome this issue, Hardt et al. [51] proposed the notion of equal opportunity (EO), which is compatible with maximal accuracy when the target label is deterministic with respect to the input features. In the probabilistic case, however, the issue is more complicated: It was shown in [46] that under differential privacy constraints, there are data sources for which EO can only be achieved at the total detriment of accuracy, in the sense that a classifier that satisfies EO cannot be more accurate than a trivial (i.e., constant) classifier. In [9] we strengthened this result by removing the privacy constraint. Namely, we have shown that for certain data sources, the most accurate classifier that satisfies EO is a trivial classifier. Furthermore, we have studied the trade-off between accuracy and EO loss (opportunity difference), and have provided a sufficient condition on the data source under which EO and non-trivial accuracy are compatible. In [19] we have further investigated the trade-off between EO difference minimization and accuracy maximization, and provided an algorithm to compute the Pareto-optimal relation between these two desiderata.

8.13 Gender and sex bias in COVID-19 epidemiological data

The COVID-19 pandemic has spurred a large amount of experimental and observational studies reporting clear correlation between the risk of developing severe COVID-19 (or dying from it) and whether the individual is male or female. In [16] we have studied the supposed male vulnerability to COVID-19 using a causal approach. We have identified a set of confounding and mediating factors, based on the review of epidemiological literature and analysis of sex-dis-aggregated data. We took those factors into consideration to produce explainable and fair prediction and decision models from observational data. The paper outlines how non-causal models can motivate discriminatory policies such as biased allocation of the limited resources in intensive care units (ICUs). The objective is to anticipate and avoid disparate impact and discrimination, by considering causal knowledge and causalbased techniques to compliment the collection and analysis of observational big-data. The hope is to contribute to more careful use of health related information access systems for developing fair and robust predictive models.

8.14 Polarization under Confirmation Bias

In our team we have developed models for polarization in multi-agent systems based on Esteban and Ray's standard family of polarization measures from economics. Agents evolve by updating their beliefs (opinions) based on an underlying influence graph, as in the standard DeGroot model for social learning, but under a confirmation bias; i.e., a discounting of opinions of agents with dissimilar views. In [12] we showed that even under this bias polarization eventually vanishes (converges to zero) if the influence graph is strongly-connected. If the influence graph is a regular symmetric circulation, we determine the unique belief value to which all agents converge. Our more insightful result in [12] establishes that, under some natural assumptions, if polarization does not eventually vanish then either there is a disconnected subgroup of agents, or some agent influences others more than she is influenced. We also proved that polarization does not necessarily vanish in weakly-connected graphs under confirmation bias. Furthermore, we showed how our model relates to the classic DeGroot model for social learning. We illustrated our model with several simulations of a running example about polarization over vaccines and of other case studies. The theoretical results and simulations in [12] provided insight into the phenomenon of polarization.

9 Bilateral contracts and grants with industry

Collaboration with the National Institute of Demographic Studies (INED)

Participants: Catuscia Palamidessi, Szilvia Lestyan, Mario Alvim, Ramon Gonze, Héber Arcolezi.

Duration: 2023–2025

Inria PI: Catuscia Palamidessi

Other partners: Universidade Federal de Minas Gerais (Brazil) and Macquarie University (Australia)

Budget for COMETE: Salary for a postdoc, working in collaboration with INED

Objectives: This project aims to study novel anonymization methods for databases published as microdata.

10 Partnerships and cooperations

10.1 International initiatives

FACTS

Participants: Frank Valencia, Mario Sergio Ferreira Alvim Junior.

Title: Foundational Approach to Cognition in Today's Society.

Program: ECOS NORD.

Duration: 2019–2023.

Coordinator: Frank Valencia.

Type of funding: The project provides funds for mobility between France and Colombia and dissemination of the results.

Other partners: Sorbonne University and Universidad Javeriana de Cali, Colombia.

Objective: This projects aims at studying the phenomenon of “Group Polarization”; the tendency for a group to learn or acquire beliefs or to make decisions that are more extreme than the initial inclinations of its members.

PROMUEVA

Participants: Frank Valencia, Carlos Pinzon Henao.

Title: Computational Models for Polarization on Social Networks Applied To Colombia Civil Unrest.

Duration: 2022–2026.

Coordinator: Frank Valencia.

Source of funding: Minciencias - Ministerio de Ciencia Tecnología e Innovación, Colombia.

Other partners: Universidad Javeriana de Cali, Colombia. Universidad del Valle, Colombia.

Objective: This projects aims at developing computational frameworks for modeling belief evolution and measuring polarization in social networks.

10.2 International research visitors

10.2.1 Visits of international scientists

Daniele Gorla

Status Associate Professor

Institution of origin: University of Rome "La Sapienza"

Country: Italy

Dates: March 2023

Context of the visit: Collaboration with Catuscia Palamidessi and Ruta Binkyte on fairness.

Mobility program/type of mobility: research stay

Martina Cinquini

Status PhD student

Institution of origin: University of Pisa

Country: Italy

Dates: September 2023 - November 2023

Context of the visit: Collaboration with Catuscia Palamidessi, Mario Alvim and Ramon Gonze on privacy and fairness.

Mobility program/type of mobility: internship

Josée Desharnais

Status Professor

Institution of origin: Laval University

Country: Canada

Dates: September 2023 - November 2023

Context of the visit: Collaboration with Catuscia Palamidessi, Mario Alvim and Ramon Gonze on privacy and fairness.

Mobility program/type of mobility: sabbatical

Annabelle McIver

Status Professor

Institution of origin: University of Macquarie

Country: Australia

Dates: December 2023

Context of the visit: Collaboration with Catuscia Palamidessi, Mario Alvim and Carroll Morgan on fairness.

Mobility program/type of mobility: research stay

Carroll Morgan

Status Professor

Institution of origin: University of New South Wales

Country: Australia

Dates: December 2023

Context of the visit: Collaboration with Catuscia Palamidessi, Mario Alvim and Annabelle McIver on fairness.

Mobility program/type of mobility: research stay

10.2.2 Visits to international teams

Frank Valencia

Visited institution Universidad Javeriana de Cali

Country: Colombia

Dates: January 2023 and July - August 2023

Context of the visit: Collaboration with Camillo Rueda and other researchers in the context of the project PROMUEVA.

Mobility program/type of mobility: research stay

10.3 European initiatives

10.3.1 Horizon Europe

ELSA

Participants: Catuscia Palamidessi, Gangsoo Zeong, Sayan Biswas.

ELSA project on cordis.europa.eu

Title: European Lighthouse on Secure and Safe AI

Duration: From September 1, 2022 to August 31, 2025

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- PAL ROBOTICS SL (PAL ROBOTICS), Spain
- YOOZ (Yooz), France
- HELSINGIN YLIOPISTO, Finland
- PLURIBUS ONE SRL, Italy
- KUNGLIGA TEKNISKA HOEGSKOLAN (KTH), Sweden
- EUROPEAN MOLECULAR BIOLOGY LABORATORY (EMBL), Germany
- THE UNIVERSITY OF BIRMINGHAM (UoB), United Kingdom
- ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE (EPFL), Switzerland
- VALEO COMFORT AND DRIVING ASSISTANCE, France
- NVIDIA SWITZERLAND AG, Switzerland
- The Alan Turing Institute, United Kingdom
- FONDAZIONE ISTITUTO ITALIANO DI TECNOLOGIA (IIT), Italy
- EIDGENOESSISCHE TECHNISCHE HOCHSCHULE ZUERICH (ETH Zürich), Switzerland
- UNIVERSITY OF LANCASTER (Lancaster University), United Kingdom
- POLITECNICO DI TORINO (POLITO), Italy
- UNIVERSITA DEGLI STUDI DI MILANO (UMIL), Italy
- CISPA - HELMHOLTZ-ZENTRUM FUR INFORMATIONSSICHERHEIT GGMBH, Germany
- LEONARDO - SOCIETA PER AZIONI (LEONARDO), Italy
- THE CHANCELLOR, MASTERS AND SCHOLARS OF THE UNIVERSITY OF OXFORD (UOXF), United Kingdom
- UNIVERSITA DEGLI STUDI DI GENOVA (UNIGE), Italy
- MAX-PLANCK-GESELLSCHAFT ZUR FORDERUNG DER WISSENSCHAFTEN EV (MPG), Germany
- CENTRE DE VISIO PER COMPUTADOR (CVC-CERCA), Spain
- UNIVERSITA DEGLI STUDI DI MODENA E REGGIO EMILIA (UNIMORE), Italy
- CONSORZIO INTERUNIVERSITARIO NAZIONALE PER L'INFORMATICA (CINI), Italy

Inria contact: Catuscia Palamedessi

Coordinator: Mario Fritz

Summary: In order to reinforce European leadership in safe and secure AI technology, we are proposing a virtual center of excellence on safe and secure AI that will address major challenges hampering the deployment of AI technology. These grand challenges are fundamental in nature. Addressing them in a sustainable manner requires a lighthouse rooted in scientific excellence and rigorous methods. We will develop a strategic research agenda which is supported by research programmes that focus on "technical robustness and safety", "privacy preserving techniques and infrastructures" and "human agency and oversight". Furthermore, we focus our efforts to detect, prevent and mitigate threats and enable recovery from harm by 3 grand challenges: "Robustness guarantees and certification", "Private and robust collaborative learning at scale" and "Human-in-the-loop decision making: Integrated governance to ensure meaningful oversight" that cut across 6 use cases: health, autonomous driving, robotics, cybersecurity, multi-media, and document intelligence. Throughout our project, we seek to integrate robust technical approaches with legal and ethical principles supported by meaningful and effective governance architectures to nurture and sustain the development and deployment of AI technology that serves and promotes foundational European values. Our initiative builds on and expands the internationally recognized, highly successful and fully operational network of excellence ELLIS (European Laboratory for Learning and Intelligent Systems). We build ELSA on its 3 pillars: research programmes, a set of research units, and a PhD/postdoc programme, thereby connecting a network of over 100 organizations and more than 337 ELLIS fellows and scholars (113 ERC grants) committed to shared standards of excellence. We will not only establish a virtual center of excellence, but all our activities will be also inclusive and open to input, interactions and collaboration of AI researchers and industrial partners in order to drive the entire field forward.

10.3.2 H2020 projects

HYPATIA

Participants: Catuscia Palamidessi, Sami Zhioua, Mario Sergio Ferreira Alvim Junior, Héber Arcolezi, Selene Leya Cerna Nahuis, Szilvia Lestyan, Sayan Biswas, Ruta Binkyte-Sadauskiene, Carlos Pinzon Henao, Karima Makhlouf.

HYPATIA project on cordis.europa.eu

Title: Privacy and Utility Allied

Duration: From October 1, 2019 to September 30, 2024

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France

Inria contact: Catuscia Palamidessi

Coordinator: Catuscia Palamidessi

Summary: With the ever-increasing use of internet-connected devices, such as computers, smart grids, IoT appliances and GPS-enabled equipments, personal data are collected in larger and larger amounts, and then stored and manipulated for the most diverse purposes. Undeniably, the big-data technology provides enormous benefits to industry, individuals and society, ranging from improving business strategies and boosting quality of service to enhancing scientific progress. On the other hand, however, the collection and manipulation of personal data raises alarming privacy issues. Both the experts and the population at large are becoming increasingly aware of the risks, due to the repeated cases of violations and leaks that keep hitting the headlines. The objective of this project is to develop the theoretical foundations, methods and tools to protect the privacy of the individuals while letting their data to be collected and used for statistical purposes. We aim in particular at developing mechanisms that: (1) can be applied and controlled directly by the user, thus avoiding the need of a trusted party, (2) are robust with respect to combination of information from different sources, and (3) provide an optimal trade-off between privacy and utility. We intend to pursue these goals by developing a new framework for privacy based on the addition of controlled noise to individual data, and associated methods to recover the useful statistical information, and to protect the quality of service.

10.3.3 Other european programs/initiatives

CRYPTECS

Title: Cloud-Ready Privacy-Preserving Technologies

Program: ANR-BMBF French-German Joint Call on Cybersecurity

Duration: 2021–2025

Coordinators: Baptiste Olivier (Orange) and Sven Trieflinger (Bosch)

Other partners: Orange (France), The Bosch Group (Germany), University of Stuttgart (Germany), Zama (SME spin-off of CryptoExperts, France), and Edgeless Systems (SME, Germany).

Inria PI: Catuscia Palamidessi

Description: The project aims at building an open-source cloud platform promoting the adoption of privacy-preserving computing (PPC) technology by offering a broad spectrum of business-ready PPC techniques (Secure Multiparty Computation, Homomorphic Encryption, Trusted Execution Environments, and methods for Statistical Disclosure Control, in particular, Differential Privacy) as reusable and composable services.

10.4 National initiatives

iPOP

Participants: Catuscia Palamidessi, Sami Zhioua, Héber Arcolezi, Sayan Biswas, Ruta Binkyte-Sadauskiene, Karima Makhlouf.

Web Page: Link

Title: Interdisciplinary Project on Privacy

Program: PEPR Cybersecurity

Duration: 1 October 2022 - 30 September 2028

Coordinators: Antoine Boutet (Insa-Lyon) - Vincent Roca (Inria)

Partners:

- Inria
- CNRS
- CNIL
- INSA-Centre Val de Loire (CVL)
- INSA-Lyon
- Université Grenoble Alpes
- Université de Lille
- Université Rennes 1
- Université de Versailles Saint-Quentin-en-Yvelines

Inria COMETE contact: Catuscia Palamidessi

Description: Digital technologies provide services that can greatly increase quality of life (e.g. connected e-health devices, location based services or personal assistants). However, these services can also raise major privacy risks, as they involve personal data, or even sensitive data. Indeed, this notion of personal data is the cornerstone of French and European regulations, since processing such data triggers a series of obligations that the data controller must abide by. This raises many multidisciplinary issues, as the challenges are not only technological, but also societal, judiciary, economic, political and ethical. The objectives of this project are thus to study the threats on privacy that have been introduced by these new services, and to conceive theoretical and technical privacy-preserving solutions that are compatible with French and European regulations, that preserve the quality of experience of the users. These solutions will be deployed and assessed, both on the technological and legal sides, and on their societal acceptability. In order to achieve these objectives, we adopt an interdisciplinary approach, bringing together many diverse fields: computer science, technology, engineering, social sciences, economy and law.

FedMalin

Participants: Catuscia Palamidessi, Sami Zhioua, Héber Arcolezi, Sayan Biswas, Ruta Binkyte-Sadauskiene, Karima Makhlouf.

Web Page: Link

Title: Federated MAchine Learning over the INternet

Program: Inria Challenge

Duration: 1 October 2022 - 30 September 2026

Coordinators: Aurélien Bellet and Giovanni Neglia

Partners:

- ARGO (Inria Paris)
- COATI (Inria Sophia)
- COMETE (Inria Saclay)
- EPIONE (Inria Sophia)
- MAGNET (Inria Lille)
- MARACAS (Inria Lyon)
- NEO (Inria Sophia)
- SPIRALS (Inria Lille)
- TRIBE (Inria Saclay)
- WIDE (Inria Rennes)

Inria COMETE contact: Catuscia Palamidessi

Description: In many use-cases of Machine Learning (ML), data is naturally decentralized: medical data is collected and stored by different hospitals, crowdsensed data is generated by personal devices, etc. Federated Learning (FL) has recently emerged as a novel paradigm where a set of entities with local datasets collaboratively train ML models while keeping their data decentralized. FedMalin aims to push FL research and concrete use-cases through a multidisciplinary consortium involving expertise in ML, distributed systems, privacy and security, networks, and medicine. We propose to address a number of challenges that arise when FL is deployed over the Internet, including privacy and fairness, energy consumption, personalization, and location/time dependencies. FedMalin will also contribute to the development of open-source tools for FL experimentation and real-world deployments, and use them for concrete applications in medicine and crowdsensing.

DIFPRIPOS

Participants: Catuscia Palamidessi.

Title: Making PostgreSQL Differentially Private for Transparent AI

Program: ANR blanc.

Duration: 2023–2026

Coordinator: Jen-François Couchot (Université de Franche-Comté).

Inria COMETE PI: Catuscia Palamidessi.

Other partners: Université de Franche-Comté, LIRIS / INSA-Lyon, The DALIBO cooperative society, and LIFO / INSA-CVL.

Objective: The general objective is to implement and to evaluate a "privacy preserving" approach for interpreting SQL queries in the sense of differential confidentiality that can be integrated into PostgreSQL.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

- Catuscia Palamidessi has co-organized and co-chaired:
  - the Fourth AAAI Workshop on Privacy-Preserving Artificial Intelligence. Washington DC, USA. February 2023.
  - the session on formal methods at the Franco-Japanese workshop on Cybersecurity. Bordeaux, France. November 29th - December 1st, 2023.
- Sami Zhioua, Catuscia Palamidessi, and other members of Comete have organized and co-chaired the Second Ethical AI @Comete workshop. Palaiseau, France. November 23-24, 2023.
- Frank Valencia has organized and chaired the Promueva Workshop on Models for Social Networks at École Polytechnique, Sorbonne Paris Nord, and IRCAM. June 12-23, 2023.
- Frank Valencia has co-organized and co-chaired the Promueva Public Workshop on Polarization in Social Networks and AI Impact, at Universidad Javeriana Cali, August 23, 2023.
- Sami Zhioua has been member of the Organizing Committee of the European Workshop on Algorithmic Fairness (EWAF'23) June 2023
- Héber Hwang Arcolezi has co-organized the 13th French Workshop on Privacy, at University Bourgogne Franche-Comté, June 12-15, 2023.

11.1.2 Scientific events: selection

- Catuscia Palamidessi is/has been program committee member of:
  - PPAI 2024. The 5th AAAI Workshop on Privacy-Preserving Artificial Intelligence
  - PETS 2024, the international conference on Privacy Enancing Technologies.
  - FOSSACS 2024, the International Conference on Foundations of Software Science and Computation Structures.

  - CSF 2024, the international IEEE Symposium on Computer Security Foundations.
  - LICS 2023, the international ACM/IEEE Conference on Logic In Computer Science.
  - CSF 2023, the international IEEE Symposium on Computer Security Foundations.
  - SDS 2023. The 10th IEEE Swiss Conference on Data Science.
  - WiL 2023. The Women in Logic Workshop.

- Frank Valencia has been program committee member of:
  - ICLP-DC 2023, the Doctoral Program International Conference on Logic Programming.
  - FOSSACS 2023, the International Conference on Foundations of Software Science and Computation Structures

- Héber Hwang Arcolezi is/has been program committee member of:
  - PETs 2024.
  - ICLR Tiny Paper Track 2024.
  - NeurIPS 2023.
  - CCS Poster Track 2023.
  - ICLR Tiny Paper Track 2023.
  - ECML / PKDD 2023.
  - FAccT 2023.
  - PPAI Workshop 2023.
  - SDS 2023.

- Sami Zhioua has been program committee member of the IEEE Afro-Mediterranean Conference on Artificial Intelligence, 2023.

11.1.3 Journal

Catuscia palamidessi is member of the editorial board of:

- (Since 2022) ACM Transactions on Privacy and Security.
- (Since 2022) TheoretiCS.
- (Since 2020) IEEE Transactions on Dependable and Secure Computing.
- (Since 2020) Journal of Logical and Algebraic Methods in Programming, Elsevier.
- (Since 2015) Acta Informatica, Springer.
- (Since 2006) Mathematical Structures in Computer Science, CUP.

11.1.4 Invited talks

- Catuscia Palamidessi has been keynote invited speaker at:
  - ACM CODASPY 2023. The 13th ACM Conference on Data and Application Security and Privacy. April 24 - 26, 2023. Charlotte, NC, United States.
  - MobiliT.AI 2023. The Annual Forum on Artificial Intelligence for critical applcations. 30-31 May 2023. MEETT Toulouse, France.
  - HYPER 2023. Workshop on Hyperproperties. 18 July 2023. Paris, France.

  - The PhD day for PhD students in ICST of the plateau de Saclay (Paris Saclay University and IP Paris). 20 june 2023. AgroParisTech, Palaiseau, France.
  - The Franco-Japanese workshop on Distributed Ledger Technologies. 14-15 November, 2023. Tokio, Japan.

- Sami Zhioua has been invited speaker at:
  - Ethical Public Robots and AI (EPURAI), Paris, France, 2023.
  - LIX École Polytechnique Monthly Seminar, Palaiseau, France, 2023.
  - Tau Team monthly seminar Université Paris-Saclay, 2023.

11.1.5 Leadership within the scientific community

Catuscia palamidessi is:

- President of SIGLOG, the ACM Special Interest Group on Logic and Computation.
- Co-chair of the of the 6th edition of the CNIL-Inria Privacy Award.
- Member of steering committees of:
  - (Since 2016) CONCUR, the International Conference in Concurrency Theory.
  - (Since 2015) EACSL, the European Association for Computer Science Logics.

11.1.6 Scientific expertise

Catuscia Palamidessi has been/is:

- Member of the Estonian Research Council for the evaluation process of the research funding applications in 2023, in the fields of Mathematics, Computer Science and Informatics.
- Member of the Romanian Research Council for the evaluation process of the research funding applications in the Innovation and Development program for 2023.
- Member of the Scientific Committee of ANR - AAPG 2024. Evaluation of project proposal in the context of Artificial Intelligence and Data Science - CES 23.
- Member of the Board of Trustees of the IMDEA Software Institute, Madrid, Spain. Since 2021.
- Member of the Sci. Adv. Board of CISPA, Helmholtz Center for Information Security. Saarbruecken, Germany. Since 2019.

11.1.7 Research administration

Catuscia palamidessi has been chair of the Scientific Committee of Inria Saclay during 2021-23.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Catuscia Palamidessi has given a Tutorial on Privacy at the GDR SRD Summer School on Distributed Learning, Lyon, France, September 2023.
- Frank Valencia has been teaching since 2019 *Concurrency Theory* and *Computability* at the Master's program of Computer Science at the University Javeriana Cali for a total of 128 hours per year.
- Héber Hwang Arcolezi was a teaching assistant for the INF361 "Introduction à l informatique" course at École Polytechnique, from April to June 2023.

- Sami Zhioua has given the following courses:
  - CSE 101 : Computer Programming I
    year: 2023-2024
    School/University: École Polytechnique
    Role: TD main instructor
  - CSE 102 : Computer Programming II
    year: 2023-2024
    School/University: École Polytechnique
    Role: TD main instructor
  - INF473X - Modal d'informatique - Cybersecurity - The Hacking Xperience
    year: 2023-2024
    School/University: École Polytechnique
    Role: TD main instructor and grader
  - Éthique dans l'apprentissage machine
    year: 2023-2024
    School/University: Aivancity
    Role: Course design and main instructor
  - CSE 103 : Introduction to Algorithms
    year: 2022-2023
    School/University: École Polytechnique
    Role: TD main instructor
  - C319: Fair Machine Learning
    year: 2022-2023
    School/University: Aivancity
    Role: Course design and main instructor (with Ruta Binkyte)

11.2.2 Supervision

Supervision of PhD students

- (2023-) Ramon Goncalves Gonze. Co-supervised by Catuscia palamidessi and Mario Alvim. Subject: Tension between privacy and utility in Census data.
- (2022-) Andreas Athanasiou. Co-supervised by Catuscia palamidessi and Kostantinos Chatzikokolakis. Subject: The shuffle model for metric differential privacy.
- (2021-) Karima Makhlouf. Co-supervised by Catuscia palamidessi and Heber Hwang Arcolezi. Subject: Relation between privacy and fairness in machine learning. Karima received the **Best Poster Award** at the workshop on Computing, Data, and Artificial Intelligence organized by the IPP doctoral schools in 2022.
- (2020-23) Ruta Binkyte-Sadauskiene. Co-supervised by Catuscia palamidessi and Sami Zhioua. Subject: Advancing Ethical AI: Methods for fairness enhancement leveraging on causality and under privacy constraints. Ruta defended her thesis in December 2023 and she is now a Postdoctoral Fellow at CISPA, Germany.

- (2020-23) Carlos Pinzon Henao. Co-supervised by Catuscia palamidessi, Frank Valencia and Pablo Piantanida. Subject: Exploring fairness and privacy in machine learning. Carlos defended his thesis in December 2023 and he is currently considering various job offers. He received the **accessit to the 2023 Doctoral Prize** awarded by the University of Paris Saclay and the Polytechnic Institute of Paris, for his work [9], which constitutes an important part of his thesis.
- (2019-23) Ganesh Del Grosso Guzman. Co-supervised by Catuscia palamidessi and Pablo Piantanida. Subject: Leakage of sensitive data from deep neural networks. Ganesh defended his thesis in November 2023 and he is now working as Researcher and Developer at Ericsson Telecommunications Inc.
- (2020-23) Sayan Biswas. Supervised by Catuscia palamidessi. Subject: Understanding and optimizing the trade-off between privacy and utility from a foundational perspective. Sayan defended his thesis in October 2023 and he is now a Postdoctoral Fellow at EPFL, Switzerland.
- (2019-23) Federica Granese. Co-supervised by Catuscia Palamidessi, Pablo Piantanida and Daniele Gorla. Subject: Towards securing machine learning algorithms. Federica defended her thesis in April 2023 and she is now a Postdoctoral Fellow in AI at IRD - UMMISCO, Sorbonne Université. Her paper [5], which constitute an important part of her thesis, was accepted as **hotspot presentation at NeurIPS 2020**. Only 5% of the paper accepted at NeurIPS have this privilege.

Supervision of postdocs and junior researchers

- (2022–2023) Szilvia Lestyan, postdoc.
- (2022–23) Selene Cerna, postdoc.
- (2022–23) Héber Hwang Arcolezi, postdoc.
- (2021–) Sami Zhioua, researcher CDD.
- (2020–) Gangsoo Zeong, junior researcher.

11.2.3 Juries

Catuscia Palamidessi has been:

- Member of the jury for the HDR defense of Benjamin Smith. October 2023.
- Member of the jury for the PhD defense of Angelo Saadeh. July 2023.
- Member of the jury for the PhD defense of Chrysoula Kosma. December 2023.
- Rapporteur and member of the jury for the PhD defense of Yakini Tchouka. December 2023.
- Rapporteur and member of the jury for the PhD defense of Sara Taki. December 2023.

11.3 Popularization

- Catuscia Palamidessi has contributed in the following dissemination actions:

  Magazine: Chut! Magazine. *Protection de la vie privée: géo-indiscernabilité.* 2023.

  Mentoring: ETAPS Mentoring Workshop. *Changing research area.* 2023.

- Frank Valencia has contributed in the following dissemination actions:

  Magazine: Chut! Magazine. *Les réseaux sociaux fournissent le milieu idéal pour la prolifération des biais cognitifs.* 2023.

Website: Inria-Saclay Communications. *Marie Spénale dessine les algorithmes des réseaux sociaux avec Inria*. 2023.

Website: Univalle Noticias. *PROMUEVA: Conociendo las claves de la polarización en redes sociales y su vínculo con la inteligencia artificial*. 2023.

TV: A Fondo de Telepacífico, Colombia. *Polarización en Redes Sociales*. 2023.

TV: Programa en Contacto, Canal Univalle. *Polarización y Redes Sociales*. 2023.

Radio: Programa Entre Líneas, Javeriana-Stereo. *Redes Sociales e Inteligencia Artificial*. 2023.

- Mario Alvim has contributed in the following dissemination actions:

Public Debate: Invited talk in the USA Census Retreat 2023 in Boston, MA, USA. Title of the talk: *How to change the mind of a government? A case study of privacy in official censuses*.

Public Debate: Invited talk at the United Nations Programme on Privacy Enhancing Technologies. 2023.

- Ruta Binkyte Sadauskiene has has contributed in the following dissemination actions:

Public Debate: Invited talk at the AIvolution debate at the Euro Parliament. *Transforming the European Union's Economy and Society*: AI Technologies, algorithms and use cases. 2023.

Online Public Debate: Women in Machine Learning and Data Science. *Ethics, image and video surveillance*. 2023.

TV: Lithuanian News Portal. *Ethical AI*. 2023.

Interview: Pause Talk (a series of interviews associated with the IEEE Returning Mothers conference). *Experience as a returning mother*. 2023.

12 Scientific production

12.1 Major publications

[1] M. S. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi and G. Smith. *The Science of Quantitative Information Flow*. Springer, 2020, pp. XXVIII, 478. DOI: 10.1007/978-3-319-96131-6. URL: https://inria.hal.science/hal-01971490.

[2] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis and C. Palamidessi. 'Geo-Indistinguishability: Differential Privacy for Location-Based Systems'. Anglais. In: *20th ACM Conference on Computer and Communications Security*. DGA, Inria large scale initiative CAPPRIS. ACM. Berlin, Allemagne: ACM Press, 2013, pp. 901–914. DOI: 10.1145/2508859.2516735. URL: http://hal.inria.fr/hal-00766821.

[3] N. E. Bordenabe, K. Chatzikokolakis and C. Palamidessi. 'Optimal Geo-Indistinguishable Mechanisms for Location Privacy'. In: Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS). Scottsdale, Arizona, United States: ACM, 2014, pp. 251–262. DOI: 10.1145/2660267.2660345. URL: https://inria.hal.science/hal-00950479.

[4] G. Cherubin, K. Chatzikokolakis and C. Palamidessi. 'F-BLEAU: Fast Black-Box Leakage Estimation'. In: *Proceedings of the 40th IEEE Symposium on Security and Privacy (SP)*. San Francisco, United States: IEEE, May 2019, pp. 835–852. DOI: 10.1109/SP.2019.00073. URL: https://hal.archives-ouvertes.fr/hal-02422945.

[5] F. Granese, M. Romanelli, D. Gorla, C. Palamidessi and P. Piantanida. 'DOCTOR: A Simple Method for Detecting Misclassification Errors'. In: Advances in Neural Information Processing Systems (NeurIPS). Proceedings. Virtual event, United States, 2021, pp. 5669–5681. URL: https://hal.science/hal-03624023.

[6] M. Guzmán, S. Haar, S. Perchy, C. Rueda and F. D. Valencia. 'Belief, Knowledge, Lies and Other Utterances in an Algebra for Space and Extrusion'. In: *Journal of Logical and Algebraic Methods in Programming* (Sept. 2016). DOI: 10.1016/j.jlamp.2016.09.001. URL: https://hal.inria.fr/hal-01257113.

[7] M. Guzmán, S. Knight, S. Quintero, S. Ramírez, C. Rueda and F. D. Valencia. 'Reasoning about Distributed Knowledge of Groups with Infinitely Many Agents'. In: *CONCUR 2019 - 30th International Conference on Concurrency Theory*. Ed. by W. Fokkink and R. van Glabbeek. Vol. 140. Amsterdam, Netherlands, Aug. 2019, 29:1–29:15. DOI: 10.4230/LIPIcs.CONCUR.2019.29. URL: https://hal.archives-ouvertes.fr/hal-02172415.

[8] S. Knight, C. Palamidessi, P. Panangaden and F. D. Valencia. 'Spatial and Epistemic Modalities in Constraint-Based Process Calculi'. In: *CONCUR 2012 - Concurrency Theory - 23rd International Conference, CONCUR 2012*. Vol. 7454. Newcastle upon Tyne, United Kingdom, Sept. 2012, pp. 317–332. DOI: 10.1007/978-3-642-32940-1. URL: http://hal.inria.fr/hal-00761116.

[9] C. Pinzón, C. Palamidessi, P. Piantanida and F. Valencia. 'On the Impossibility of non-Trivial Accuracy in Presence of Fairness Constraints'. In: Proceedings of the AAAI 36th Conference on Artificial Intelligence. Vol. 36. Proceedings 7. Vancouver / Virtual, Canada, 30th June 2022, pp. 7993–8000. DOI: 10.1609/aaai.v36i7.20770. URL: https://hal.science/hal-03452324.

[10] M. Romanelli, K. Chatzikokolakis, C. Palamidessi and P. Piantanida. 'Estimating g-Leakage via Machine Learning'. In: *CCS '20 - 2020 ACM SIGSAC Conference on Computer and Communications Security*. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS). Online, United States: ACM, 9th Nov. 2020, pp. 697–716. URL: https://hal.science/hal-03091469.

12.2 Publications of the year

International journals

[11] G. Alves, F. Bernier, M. Couceiro, K. Makhlouf, C. Palamidessi and S. Zhioua. 'Survey on Fairness Notions and Related Tensions'. In: *EURO journal on decision processes* (2023). DOI: 10.1016/j.ejdp.2023.100033. URL: https://hal.science/hal-03484009.

[12] M. S. Alvim, B. Amorim, S. Knight, S. Quintero and F. Valencia. 'A Formal Model for Polarization under Confirmation Bias in Social Networks'. In: *Logical Methods in Computer Science* (7th Mar. 2023). DOI: 10.46298/lmcs-19(1:18)2023. URL: https://hal.science/hal-03872692.

[13] H. H. Arcolezi, S. Gambs, J.-F. Couchot and C. Palamidessi. 'On the Risks of Collecting Multidimensional Data Under Local Differential Privacy'. In: *Proceedings of the VLDB Endowment (PVLDB)* 16.5 (Jan. 2023), pp. 1126–1139. DOI: 10.14778/3579075.3579086. URL: https://inria.hal.science/hal-04082592.

[14] S. Biswas and C. Palamidessi. 'PRIVIC: A privacy-preserving method for incremental collection of location data'. In: *Proceedings on Privacy Enhancing Technologies* 2024.1 (23rd Oct. 2023), pp. 582–596. DOI: 10.56553/popets-2024-0033. URL: https://inria.hal.science/hal-03968692.

[15] G. Del Grosso, G. Pichler, C. Palamidessi and P. Piantanida. 'Bounding Information Leakage in Machine Learning'. In: *Neurocomputing* 534 (2023), pp. 1–17. DOI: 10.1016/j.neucom.2023.02.058. URL: https://inria.hal.science/hal-04349219.

[16] N. Díaz-Rodríguez, R. Binkytė, W. Bakkali, S. Bookseller, P. Tubaro, A. Bacevičius, S. Zhioua and R. Chatila. 'Gender and sex bias in COVID-19 epidemiological data through the lens of causality'. In: *Information Processing and Management* 60.3 (May 2023), p. 103276. DOI: 10.1016/j.ipm.2023.103276. URL: https://hal.science/hal-03961804.

[17] N. Fernandes, A. Mciver, C. Palamidessi and M. Ding. 'Universal optimality and robust utility bounds for metric differential privacy'. In: *Journal of Computer Security* (18th July 2023), pp. 1–42. DOI: 10.3233/JCS-230036. URL: https://inria.hal.science/hal-04349262.

[18] F. Galli, K. Jung, S. Biswas, C. Palamidessi and T. Cucinotta. 'Advancing Personalized Federated Learning: Group Privacy, Fairness, and Beyond'. In: *SN Computer Science*. Volume 4, issue 6, November 2023 4.6 (28th Oct. 2023), p. 831. DOI: 10.1007/s42979-023-02292-0. URL: https://hal.science/hal-04320177.

[19] C. Pinzón, C. Palamidessi, P. Piantanida and F. Valencia. 'On the incompatibility of accuracy and equal opportunity'. In: *Machine Learning* (2nd May 2023). DOI: 10.1007/s10994-023-06331-y. URL: https://hal.science/hal-04308195.

Invited conferences

[20] C. Palamidessi. 'Local Methods for Privacy Protection and Impact on Fairness'. In: CODASPY 2023 - Thirteenth ACM Conference on Data and Application Security and Privacy. Charlotte NC, United States: ACM, 24th Apr. 2023. DOI: 10.1145/3577923.3587263. URL: https://inria.hal.science/hal-04349271.

International peer-reviewed conferences

[21] M. S. Alvim, N. Fernandes, B. D. Nogueira, C. Palamidessi and T. V. A. Silva. 'On the Duality of Privacy and Fairness'. In: *International Conference on AI and the Digital Economy (CADE 2023),* CADE 2023 - International Conference on AI and the Digital Economy. Venice, Italy, 26th June 2023, p. 46 –48. URL: https://hal.science/hal-04407491.

[22] H. H. Arcolezi, S. Cerna and C. Palamidessi. 'On the Utility Gain of Iterative Bayesian Update for Locally Differentially Private Mechanisms'. In: *Lecture Notes in Computer Science*. DBSec 2023 - 37th IFIP Annual Conference on Data and Applications Security and Privacy. Vol. LNCS-13942. Data and Applications Security and Privacy XXXVII. Sophia Antipolis, France: Springer Nature Switzerland, 12th July 2023, pp. 165–183. DOI: 10.1007/978-3-031-37586-6_11. URL: https://inria.hal.science/hal-04175035.

[23] H. H. Arcolezi, K. Makhlouf and C. Palamidessi. '(Local) Differential Privacy has NO Disparate Impact on Fairness'. In: *Lecture Notes in Computer Science*. DBSec 2023 - 37th IFIP Annual Conference on Data and Applications Security and Privacy. Vol. LNCS-13942. Data and Applications Security and Privacy XXXVII. SOPHIA ANTIPOLIS, France: Springer Nature Switzerland, 12th July 2023, pp. 3–21. DOI: 10.1007/978-3-031-37586-6_1. URL: https://inria.hal.science/hal-04175027.

[24] H. H. Arcolezi, C. Palamidessi, C. Pinzón and S. Gambs. 'Frequency Estimation of Evolving Data Under Local Differential Privacy'. In: EDBT 2023 - 26th International Conference on Extending Database Technology. Ioánnina, Greece, 28th May 2023, pp. 512–525. DOI: 10.48786/edbt.2023.44. URL: https://inria.hal.science/hal-03911550.

[25] S. Cerna and C. Palamidessi. 'On the Application and Impact of differential privacy and Fairness in Ambulance Engagement Time Prediction'. In: *Proceedings of the Tiny Papers Track at ICLR*. ICLR 2023 - The First Tiny Papers Track at ICLR 2023. Proceedings of the Tiny Papers Track at ICLR. Kigali, Rwanda, 2023. URL: https://inria.hal.science/hal-04349309.

[26] K. Chatzikokolakis, G. Cherubin, C. Palamidessi and C. Troncoso. 'Bayes Security: A Not So Average Metric'. In: *Proceedings of the IEEE 36th Computer Security Foundations Symposium (CSF)*. CSF 2023 - 36th IEEE Computer Security Foundations Symposium. Proceedings of the IEEE 36th Computer Security Foundations Symposium (CSF). Dubrovnik, Croatia: IEEE, 2023. DOI: 10.1109/CSF57540.2023.00011. URL: https://inria.hal.science/hal-04349285.

[27] F. Galli, S. Biswas, K. Jung, T. Cucinotta and C. Palamidessi. 'Group privacy for personalized federated learning'. In: Proceedings of the 9th International Conference on Information Systems Security and Privacy - ICISSP. Lisbon, Portugal: SCITEPRESS - Science and Technology Publications, 2023, pp. 252–263. DOI: 10.5220/0011885000003405. URL: https://inria.hal.science/hal-03907130.

[28] D. Gorla, L. Jalouzot, F. Granese, C. Palamidessi and P. Piantanida. 'On the (Im)Possibility of Estimating Various Notions of Differential Privacy (short paper)'. In: *Proceedings of the 24th Italian Conference on Theoretical Computer Science*. ICTCS 2023 - The 24th Italian Conference on Theoretical Computer Science. Vol. 3587. Proceedings of the 24th Italian Conference on Theoretical Computer Science. Palermo, Italy, 2023, pp. 219–224. URL: https://inria.hal.science/hal-04349303.

[29] M. Jurado, R. G. Gonze, M. S. Alvim and C. Palamidessi. 'Analyzing the Shuffle Model Through the Lens of Quantitative Information Flow'. In: *Proceedings of the IEEE 36th Computer Security Foundations Symposium (CSF)*. CSF 2023 - 36th IEEE Computer Security Foundations Symposium. Proceedings of the IEEE 36th Computer Security Foundations Symposium (CSF). Dubrovnik, Croatia: IEEE, 22nd May 2023, pp. 423–438. DOI: 10.1109/CSF57540.2023.00033. URL: https://inria.hal.science/hal-04349295.

[30] S. Simon, C. Petrui, C. Pinzón and C. Palamidessi. 'Obfuscation Padding Schemes that Minimize Rényi Min-Entropy for Privacy'. In: *Lecture Notes in Computer Science*. ISPEC 2023 - The 18th International Conference on Information Security Practice and Experience. Vol. LNCS-14341. Information Security Practice and Experience. Coppenhagen, Denmark: Springer Nature Singapore, 8th Nov. 2023, pp. 74–90. DOI: 10.1007/978-981-99-7032-2_5. URL: https://hal.science/hal-04322523.

Doctoral dissertations and habilitation theses

[31] R. Binkytė. 'Advancing Ethical AI: Methods for fairness enhancement leveraging on causality and under privacy constraints'. Ecole Polytechnique (EDX), 19th Dec. 2023. URL: https://hal.science/tel-04407125.

[32] S. Biswas. 'Understanding and optimizing the trade-off between privacy and utility from a foundational perspective'. Ecole Polytechnique (EDX), 18th Oct. 2023. URL: https://hal.science/tel-04407120.

[33] G. Del Grosso. 'Leakage of Sensitive Data from Deep Neural Networks'. Ecole Polytechnique (EDX), 7th Nov. 2023. URL: https://hal.science/tel-04407131.

[34] F. Granese. 'Towards Securing Machine Learning Algorithms through Misclassification Detection and Adversarial Attack Detection'. Ecole Polytechnique (EDX); Sapienza University of Rome, 21st Apr. 2023. URL: https://hal.science/tel-04407139.

[35] C. Pinzón. 'Exploring fairness and privacy in machine learning'. Ecole Polytechnique, 6th Dec. 2023. URL: https://hal.science/tel-04407152.

Reports & preprints

[36] R. Binkytė, L. Grozdanovski and S. Zhioua. *On the Need and Applicability of Causality for Fair Machine Learning*. 1st Nov. 2023. URL: https://inria.hal.science/hal-04329115.

[37] R. Binkytė, S. Zhioua and Y. Turki. *Dissecting Causal Biases*. 21st Jan. 2024. URL: https://inria.hal.science/hal-04329098.

[38] K. Makhlouf, H. Hwang Arcolezi, S. Zhioua, G. B. Brahim and C. Palamidessi. *On the Impact of Multi-dimensional Local Differential Privacy on Fairness*. 7th Dec. 2023. URL: https://hal.science/hal-04329938.

[39] R. Panainte, Y. Turki and S. Zhioua. *A Web Application Software for Causal-based Machine Learning Discrimination Estimation*. 12th Feb. 2024. URL: https://inria.hal.science/hal-04355882.

[40] C. Pinzón and K. Jung. *Fast Python sampler for the von Mises Fisher distribution.* 3rd Aug. 2023. URL: https://hal.science/hal-04004568.

[41] S. Zhioua and R. Binkytė. *Shedding light on underrepresentation and Sampling Bias in machine learning.* 7th Dec. 2023. URL: https://inria.hal.science/hal-04329092.

12.3 Cited publications

[42] M. S. Alvim, K. Chatzikokolakis, Y. Kawamoto and C. Palamidessi. 'Information Leakage Games: Exploring Information as a Utility Function'. In: *ACM Transactions on Privacy and Security* 25.3 (2022). Journal version of GameSec'17 paper (arXiv:1705.05030). DOI: 10.1145/3517330. URL: https://hal.science/hal-03091413.

[43] M. S. Alvim, K. Chatzikokolakis, C. Palamidessi and G. Smith. 'Measuring Information Leakage Using Generalized Gain Functions'. In: *Proceedings of the 25th IEEE Computer Security Foundations Symposium (CSF).* 2012, pp. 265–279. DOI: 10.1109/CSF.2012.26. URL: http://hal.inria.fr/hal-00734044/en.

[44] H. Brenner and K. Nissim. 'Impossibility of Differentially Private Universally Optimal Mechanisms'. In: *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS).* IEEE Computer Society, Oct. 2010, pp. 71–80. URL: http://doi.ieeecomputersociety.org/10.1109/FOCS.2010.13.

[45] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe and C. Palamidessi. 'Broadening the scope of Differential Privacy using metrics'. In: *Proceedings of the 13th International Symposium on Privacy Enhancing Technologies (PETS 2013).* Ed. by E. De Cristofaro and M. Wright. Vol. 7981. Lecture Notes in Computer Science. Springer, 2013, pp. 82–102. URL: https://inria.hal.science/hal-00767210.

[46] R. Cummings, V. Gupta, D. Kimpara and J. Morgenstern. 'On the Compatibility of Privacy and Fairness'. In: *Proceedings of the 27th Conference on User Modeling, Adaptation and Personalization.* UMAP'19 Adjunct. Larnaca, Cyprus: Association for Computing Machinery, 2019, pp. 309–315. DOI: 10.1145/3314183.3323847. URL: https://doi.org/10.1145/3314183.3323847.

[47] M. D. Ekstrand, R. Joshaghani and H. Mehrpouyan. 'Privacy for All: Ensuring Fair and Equitable Privacy Protections'. In: *Proceedings of the First ACM Conference on Fairness, Accountability and Transparency (FAT).* Ed. by S. A. Friedler and C. Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 2018, pp. 35–47. URL: http://proceedings.mlr.press/v81/ekstrand18a.html.

[48] J.-M. Esteban and D. Ray. 'On the Measurement of Polarization'. In: *Econometrica* 62.4 (1994), pp. 819–851. URL: http://www.jstor.org/stable/2951734.

[49] A. Ghosh, T. Roughgarden and M. Sundararajan. 'Universally utility-maximizing privacy mechanisms'. In: *Proceedings of the 41st annual ACM Symposium on Theory of Computing (STOC).* Bethesda, MD, USA: ACM, 2009, pp. 351–360. DOI: http://doi.acm.org/10.1145/1536414.1536464. URL: http://doi.acm.org/10.1145/1536414.1536464.

[50] F. Granese, D. Gorla and C. Palamidessi. 'Enhanced Models for Privacy and Utility in Continuous-Time Diffusion Networks'. In: *International Journal of Information Security* 20.5 (2021), pp. 673–782. DOI: 10.1007/s10207-020-00530-7. URL: https://hal.inria.fr/hal-03094843.

[51] M. Hardt, E. Price and N. Srebro. 'Equality of Opportunity in Supervised Learning'. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS).* NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 3323–3331.

[52] J. Jia, A. Salem, M. Backes, Y. Zhang and N. Z. Gong. 'MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples'. In: *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS).* CCS '19. London, United Kingdom: Association for Computing Machinery, 2019, pp. 259–274. DOI: 10.1145/3319535.3363201. URL: https://doi.org/10.1145/3319535.3363201.

[53] M. Romanelli, K. Chatzikokolakis and C. Palamidessi. 'Optimal Obfuscation Mechanisms via Machine Learning'. In: *CSF 2020 - 33rd IEEE Computer Security Foundations Symposium*. Preprint version of a paper that appeared on the Proceedings of the IEEE 33rd Computer Security Foundations Symposium, CSF 2020. Online, United States: IEEE, June 2020, pp. 153–168. URL: https://hal.inria.fr/hal-03091514.

[54] L. Song, R. Shokri and P. Mittal. 'Privacy Risks of Securing Machine Learning Models against Adversarial Examples'. In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*. Ed. by L. Cavallaro, J. Kinder, X. Wang and J. Katz. ACM, 2019, pp. 241–257. DOI: 10.1145/3319535.3354211. URL: https://doi.org/10.1145/3319535.3354211.

[55] M. C. Tschantz, S. Sen and A. Datta. 'SoK: Differential Privacy as a Causal Property'. In: *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*. IEEE, 2020, pp. 354–371. DOI: 10.1109/SP40000.2020.00012. URL: https://doi.org/10.1109/SP40000.2020.00012.