RESEARCH CENTRE

Inria Saclay Centre at Institut Polytechnique de Paris

IN PARTNERSHIP WITH:

CNRS, Institut Polytechnique de Paris

2024

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

Inria

Contents

Project-Team COMETE

Creation of the Project-Team: 2021 December 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]
- Frank Valencia [CNRS, Researcher]
- Sami Zhioua [LIX, Researcher, until Aug 2024]

Post-Doctoral Fellow

- Carlos Pinzon Henao [INRIA, Post-Doctoral Fellow, from Jul 2024]

PhD Students

- Andreas Athanasiou [INRIA]
- Ramon Goncalves Gonze [INRIA]
- Karima Makhlouf [INRIA, until Oct 2024]

Technical Staff

- Ehab ElSalamouny [INRIA, Engineer, from Nov 2024]
- Gangsoo Zeong [INRIA, Engineer]

Interns and Apprentices

- Ranim Bouzamoucha [INRIA, Intern, from May 2024 until Jul 2024]
- Ayoub Ouni [INRIA, Intern, from May 2024 until Jul 2024]
- Mohamed Rejili [INRIA, Intern, from Jun 2024 until Aug 2024]
- Ahmed Semah Zouaghi [INRIA, Intern, from Mar 2024 until May 2024]

Administrative Assistant

- Mariana De Almeida [INRIA]

Visiting Scientists

- Fabio Gadducci [UNIV PISE, from Mar 2024 until Apr 2024]
- Artur Gaspar Da Silva [UFMG, until Feb 2024]
- Annabelle-Kate Mciver [UNIV MACQUARIE, until Jan 2024]
- Charles-Carroll Morgan [USYD, until Jan 2024]

External Collaborators

- Sayan Biswas [EPFL - Lausanne, from May 2024]
- Konstantinos Chatzikokolakis [CNRS]
- Mario Sergio Ferreira Alvim Junior [UFMG, until Mar 2024]
- Szilvia Lestyan [INED]
- Judith Sainz-Pardo Diaz [CSIC, from Sep 2024]

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 [35]. 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:

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

- 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 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 [3] (cfr. Figure 2) with mechanisms that allow to take into

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).

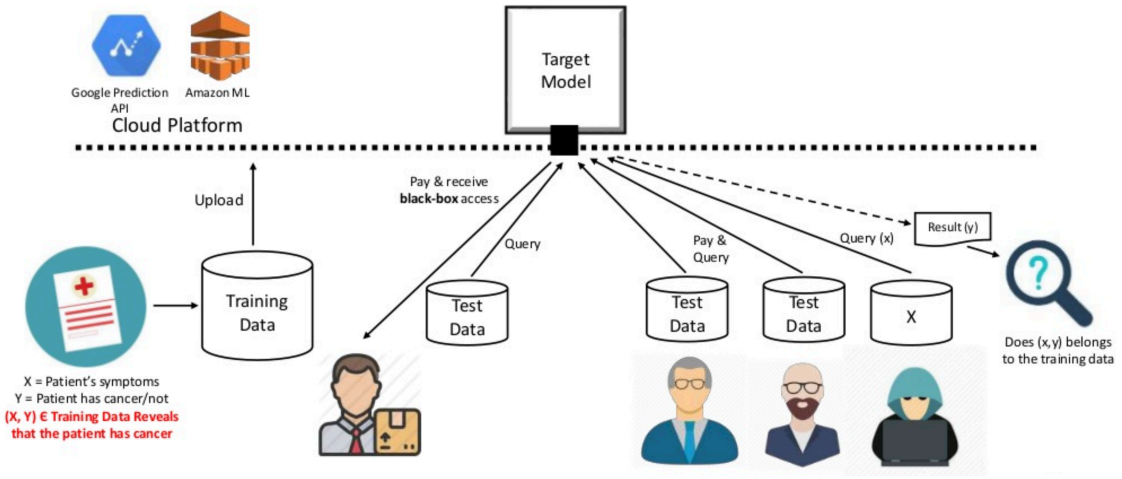

Figure 3: Privacy breach in machine learning as a service.

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 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. [41]). On the other hand, there are studies that show that there is a trade-off between protection from inference attacks and robustness [43]. 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* [37] and of *Equality of False Positives* and *Equality of False Negatives* [36]. 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* [44].

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 [33] 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* [32] 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 [11]. 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 [42]. 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 [39] 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 [38]. 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 HDR

Sami Zhioua got his HDR in July 2024.

6.2 Awards

Test of Time Award at the International IEEE Conference on Computer Security Foundations (CSF 2024). It was given for the paper "Measuring Information Leakage using Generalized Gain Functions", by Kostas Chatzikokolakis, Mario Alvim, Catuscia Palamidessi and Geoffrey Smith, that was published at CSF 2012.

Frank Valencia's paper On "Fairness and Consensus in an Asynchronous Opinion Model for Social Networks" was nominated to Best Paper at CONCUR 2024.

6.3 New Projects

Frank Valencia is the principal investigator of a new interdisciplinary project funded by CNRS-MITI, titled *Testing Opinion Biases in Social Networks (TOBIAS)*. This two-year project is

conducted in collaboration with Jean-Claude Dreher from the *Cognitive Neuroscience Centre-UMR 5229*. TOBIAS aims to validate the models developed in the PROMUEVA project through behavioral experiments and functional Magnetic Resonance Imaging (fMRI) studies, investigating how individuals integrate information in social contexts and the neural mechanisms underlying these processes.

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, Sébastien 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, Sébastien 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, Sébastien 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 Sergio 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

7.1.10 libqif - A Quantitative Information Flow C++ Toolkit Library

Keywords: Information leakage, Privacy, C++, Linear optimization

Functional Description: The goal of libqif is to provide an efficient C++ toolkit implementing a variety of techniques and algorithms from the area of quantitative information flow and differential privacy. We plan to implement all techniques produced by Com\'ete in recent years, as well as several ones produced outside the group, giving the ability to privacy researchers to reproduce our results and compare different techniques in a uniform and efficient framework.

Some of these techniques were previously implemented in an ad-hoc fashion, in small, incompatible with each-other, non-maintained and usually inefficient tools, used only for the purposes of a single paper and then abandoned. We aim at reimplementing those – as well as adding several new ones not previously implemented – in a structured, efficient and maintainable manner, providing a tool of great value for future research. Of particular interest is the ability to easily re-run evaluations, experiments, and case-studies from QIF papers, which will be of great value for comparing new research results in the future.

The library's development continued in 2020 with several new added features. 68 new commits were pushed to the project's git repository during this year. The new functionality was directly applied to the experimental results of several publications of COMETE.

URL: https://github.com/chatziko/libqif

Contact: Konstantinos Chatzikokolakis

7.1.11 IBU: A java library for estimating distributions

Keywords: Privacy, Statistic analysis, Bayesian estimation

Functional Description: The main objective of this library is to provide an experimental framework for evaluating statistical properties on data that have been sanitized by obfuscation mechanisms, and for measuring the quality of the estimation. More precisely, it allows modeling the sensitive data, obfuscating these data using a variety of privacy mechanisms, estimating the probability distribution on the original data using different estimation methods, and measuring the statistical distance and the Kantorovich distance between the original and estimated distributions. This is one of the main software projects of Palamidessi's ERC Project HYPATIA.

We intend to extend the software with functionalities that will allow estimating statistical properties of multi-dimensional (locally sanitized) data and using collections of data locally sanitized with different mechanisms.

URL: https://gitlab.com/locpriv/ibu

Contact: Ehab Elsalamouny

7.1.12 ldp-audit

Name: Local Differential Privacy Auditor

Keyword: Differential privacy

Functional Description: A tool for auditing Locally Differentially Private (LDP) protocols.

URL: https://github.com/hharcolezi/ldp-audit

Contact: Heber Hwang Arcolezi

7.1.13 Polarizómetro

Name: Polarizómetro

Keyword: Social networks

Functional Description: The Polarizómetro is a platform that was launched in August 2024 in a public event (https://sites.google.com/view/promueva/eventos/2024) with an audience of about 200 people. This platform, meant for decision-makers and available online, allows to measure the polarization of an opinion distribution in a group or social media over a particular subject. The opinion can be expressed as usual posts on social media or a standard Likert scale. The polarization can be measured using several standard notions from the literature such as Esteban and Ray's, or using our measure MEC (the Minimal Effort to Consensus) developed in our project PROMUEVA based on the Earth Mover Distance.

The platform has been used to regularly measure polarization on real opinion distributions in the social media X (formerly known as Twitter) about the Pension Reform in Colombia and about the benefits of the 2024 United Nations Biodiversity Conference of the Parties (COP16) that took place in Cali, Colombia.

URL: https://polarizometro.lipn.univ-paris13.fr/

Contact: Frank Valencia

Partners: LIPN (Laboratoire d'Informatique de l'Université Paris Nord), Pontificia Universidad Javeriana Cali

8 New results

Participants: Catuscia Palamidessi, Frank Valencia, Sami Zhioua, Heber Hwang Arcolezi, Gangsoo Zeong, Sayan Biswas, Ruta Binkyte-Sadauskiene, Ramon Gonze, Szilvia Lestyan, Karima Makhlouf, Carlos Pinzon Henao, Andreas Athanasiou.

8.1 Privacy

8.1.1 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), [3]. 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.

Electric vehicles (EVs) are becoming more popular due to environmental consciousness. The limited availability of charging stations (CSs), compared to the number of EVs on the road, has led to increased range anxiety and a higher frequency of CS queries during trips. Simultaneously, personal data use for analytics is growing at an unprecedented rate, raising concerns for privacy. One standard for formalising location privacy is the geo-indistinguishability framework mentioned above. However, the noise must be tuned properly, considering the implications of potential utility losses. In [13], we have introduced the notion of approximate geo-indistinguishability (AGeoI), which allows EVs to obfuscate their query locations while remaining within their area of interest. It is vital because journeys are often sensitive to a sharp drop in quality of service (QoS). Our method applies AGeoI with dummy data generation to provide two-fold privacy protection for EVs while preserving a high QoS. Analytical insights and experiments demonstrate that the majority of EVs get “privacy-for-free” and that the utility loss caused by the gain in privacy guarantees is minuscule. In addition to providing high QoS, the iterative Bayesian update allows for a private and precise CS occupancy forecast, which is crucial for unforeseen traffic congestion and efficient route planning.

8.1.2 Improving the utility of local differential privacy

We have investigated the problem of collecting multidimensional data throughout time (i.e., longitudinal studies) for the fundamental task of frequency estimation under Local Differential Privacy (LDP) guarantees. Contrary to frequency estimation of a single attribute, the multidimensional aspect demands particular attention to the privacy budget. Besides, when collecting user statistics longitudinally, privacy progressively degrades. Indeed, the "multiple" settings in combination (i.e., many attributes and several collections throughout time) impose several challenges, for which this paper proposes the first solution for frequency estimates under LDP. To tackle these issues, in [12] we have extended the analysis of three state-of-the-art LDP protocols (Generalized Randomized Response-GRR, Optimized Unary Encoding-OUE, and Symmetric Unary Encoding-SUE) for both longitudinal and multidimensional data collections. While the known literature uses OUE and SUE for two rounds of sanitization (a.k.a. memoization), i.e., L-OUE and L-SUE, respectively, we have analytically and experimentally shown that starting with OUE and then with SUE provides higher data utility (i.e., L-OSUE). Also, for attributes with small domain sizes, we have proposed Longitudinal GRR (L-GRR), which provides higher utility than the other protocols based on unary encoding. Last, we have also propose a new solution named Adaptive LDP for LOngitudinal and Multidimensional FREquency Estimates (ALLOMFREE), which randomly samples a single attribute to be sent with the whole privacy budget and adaptively selects the optimal protocol, i.e., either L-GRR or L-OSUE. As shown in the results, ALLOMFREE consistently and considerably outperforms the state-of-the-art L-SUE and L-OUE protocols in the quality of the frequency estimates.

8.1.3 Auditing local differential privacy

While the existing literature on Differential Privacy (DP) auditing predominantly focuses on the centralized model (e.g., in auditing the DP-SGD algorithm), we advocate for extending this approach to audit Local DP (LDP). To achieve this, in [15] we have introduced the LDP-Auditor framework for empirically estimating the privacy loss of locally differentially private mechanisms. This approach leverages recent advances in designing privacy attacks against LDP frequency estimation protocols. More precisely, in this paper, through the analysis of numerous state-of-the-art LDP protocols, we extensively explore the factors influencing the privacy audit, such as the impact of different encoding and perturbation functions. Additionally, we investigate the influence of the domain size and the theoretical privacy loss parameters ϵ and δ on local privacy estimation. In-depth case studies are also conducted to explore specific aspects of LDP auditing, including distinguishability attacks on LDP protocols for longitudinal studies and multidimensional data. Finally, we present a notable achievement of our LDP-Auditor framework, which is the discovery of a bug in a state-of-the-art LDP Python package. Overall, our LDP-Auditor framework as well as our study offer valuable insights into the sources of randomness and information loss in LDP protocols. These contributions collectively provide a realistic understanding of the local privacy loss, which can help practitioners in selecting the LDP mechanism and privacy parameters that best align with their specific requirements. We open-sourced LDP-Auditor in github.com/hharcolezi/ldp-audit (BIL entry).

8.1.4 Privacy in machine learning

Training differentially private machine learning models requires constraining an individual's contribution to the optimization process. This is achieved by clipping the 2-norm of their gradient at a predetermined threshold prior to averaging and batch sanitization. This selection adversely influences optimization in two opposing ways: it either exacerbates the bias due to excessive clipping at lower values, or augments sanitization noise at higher values. The choice significantly hinges on factors such as the dataset, model architecture, and even varies within the same optimization, demanding meticulous tuning usually accomplished through a grid search. In order to circumvent the privacy expenses incurred in hyperparameter tuning, in [24] we have proposed a novel approach to dynamically optimize the clipping threshold. We treat this threshold as an additional learnable parameter, establishing a clean relationship between the threshold and the cost function. This allows us to optimize the former with gradient descent, with minimal repercussions on the overall privacy analysis. Our method is thoroughly assessed against alternative fixed and adaptive strategies across diverse datasets, tasks, model dimensions, and privacy levels. Our results indicate that it performs comparably or better in the evaluated scenarios, given the same privacy requirements.

Federated Learning (FL) enables clients to train a joint model without disclosing their local data. Instead, they share their local model updates with a central server that moderates the process and creates a joint model. However, FL is susceptible to a series of privacy attacks. Recently, the source inference attack (SIA) has been proposed where an honest-but-curious central server tries to identify exactly which client owns a specific data record. In [21], we have proposed a defense against SIAs by using a trusted shuffler, without compromising the accuracy of the joint model. We employ a combination of unary encoding with shuffling, which can effectively blend all clients' model updates, preventing the central server from inferring information about each client's model update separately. In order to address the increased communication cost of unary encoding we employ quantization. Our preliminary experiments show promising results; the proposed mechanism notably decreases the accuracy of SIAs without compromising the accuracy of the joint model.

8.2 Fairness

8.2.1 Relation between causality, fairness and privacy

Addressing the problem of fairness is crucial to safely using machine learning algorithms to support decisions that have 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. However, the most recent notions of fairness are causal-based and reflect the now widely accepted idea that using causality is necessary to appropriately address the problem of fairness. In [17] we have examined an exhaustive list of causal-based fairness notions and studied their applicability in real-world scenarios. As most causal-based fairness notions are defined in terms of non-observable quantities (e.g., interventions and counterfactuals), their deployment in practice requires computing or estimating those quantities using observational data. Our paper offers a comprehensive report of the different approaches to infer causal quantities from observational data, including identifiability (Pearl's SCM framework) and estimation (potential outcome framework). The main contributions of our survey paper are (1) a guideline to help select a suitable causal fairness notion given a specific real-world scenario and (2) a ranking of the fairness notions according to Pearl's causation ladder, indicating how difficult it is to deploy each notion in practice.

Local differential privacy is based on the application of controlled noise on the data. The introduction of noise, however, inevitably affects the utility of the data, particularly by distorting the correlations between individual data components. This distortion can prove detrimental to tasks such as causal structure learning. In [23], we have considered various well-known locally differentially private mechanisms and compared the trade-off between the privacy they provide, and the accuracy of the causal structure produced by algorithms for causal learning when applied to data obfuscated by these mechanisms. Our analysis yields valuable insights for selecting appropriate local differentially private protocols for causal discovery tasks. We foresee that our findings will aid researchers and practitioners in conducting locally private causal discovery.

8.2.2 Relation between local differential privacy and fairness

One main line of research that we have pursued recently is the study of the interaction between fairness and privacy in the context of machine learning. This work has lead to various publications: In [25] we have developed the first formal analysis of the effect of local differential privacy (LDP) on fairness. More precisely, we have considered two ethical problems in machine learning: the possible leakage of sensitive information about the training data, and the possible unfair predictions of a model trained on biased data. As both these issues are critical for an AI respectful of fundamental human rights, it is necessary to address them at the same time, and it is therefore crucial o understand how they interact. we have focused on the so-called k-random response (k-RR) mechanism, motivated by the fact that it is the most representative for LDP, in the sense that it performs the exact amount of obfuscation that it is needed to achieve LDP. Furthermore, k-RR is the basis of real life applications, such as Google's RAPPOR. Concerning fairness, we have considered some of the most popular notions of fairness: statistical parity, conditional statistical parity, and equal opportunities. In this context, we have been able to derive precise mathematical formulas describing how the amount of (deviation from) fairness in the data can be affected by the application of k-RR to them, and how this effect depends on ϵ, the privacy parameter of k-RR.

We would like to underline the importance of such formal analysis: there have been a lot of studies on the connection between privacy and fairness in recent years, but they were all experimental, and, interestingly, they were reaching opposite conclusions, in the sense that some of them were showing that LDP tends to remove unfairness, while others were showing the contrary. Our work shows exactly under which conditions, and in what sense, LDP affects fairness. Furthermore, our formal analysis allowed us to derive surprising phaenomena, like the fact that, in certain cases, LPD can not only remove the bias against a certain group, but even steer the bias in the opposite direction. Namely, under LDP, the un-privileged group may become the privileged one.

In addition to the theoretical analyses, we has also conducted experimental studies on the impact of LPD on fairness and utility (accuracy, F1, AUC and recall) in the context of machine learning, contributions that are contained in the second and third papers. In particular, in [16] we have conducted studies on the effect of LDP on multi-dimensional data, comparing different techniques for their obfuscation (e.g., applying noise to each attribute independently, or in a combined fashion). An important observation resulting from these empirical studies is that these techniques effectively reduce disparity, and that the different techniques diverge in their performances only at low privacy guarantees. Another important conclusion is that the distribution on the decision has an important

effect on which group is more sensitive to the obfuscation, thus confirming the theoretical studies of [25].

In [34] we have conducted an empirically study of how the collection of multiple sensitive attributes can impact fairness, and the role of LPD in this impact. In particular, we have proposed a novel privacy budget allocation scheme that generally led to a better privacy-utility-fairness trade-off 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. This empirical study motivated the theoretical studies of [25], and challenged the (until then) common belief that differential privacy necessarily leads to worsened fairness in machine learning. [34] was recipient of the Best Paper Award at DBSEC 2023.

8.2.3 Relation between 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. [40] 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 [36] 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 [10] 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 [18] 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.

8.2.4 A Bayesian approach to eliminate bias from training data

In [23], we have considered the problem of unfair discrimination between two groups and proposed a pre-processing method to achieve fairness. Corrective methods like statistical parity usually lead to bad accuracy and do not really achieve fairness in situations where there is a correlation between the sensitive attribute S and the legitimate attribute E (explanatory variable) that should determine the decision. To overcome these drawbacks, other notions of fairness have been proposed, in particular, conditional statistical parity and equal opportunity, that prescribe that any difference in the prediction between the two group should be justified by E or by the "true decision" based on E. However, E is often not directly observable in the data. We may observe some other variable Z representing E, but the problem is that Z may also be affected by S, hence Z itself can be biased. To deal with this problem, in [22] we have proposed BaBE (Bayesian Bias Elimination), an approach based on a combination of Bayes inference and the Expectation-Maximization method, to estimate the most likely value of E for a given Z for each group. The decision can then be based directly on the estimated E. We show, by experiments on synthetic and real data sets, that our approach provides a good level of fairness as well as high accuracy.

8.3 Models for polarization in social networks

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 studied these dynamics in a model that uses a *influence graph* i.e., a weighted directed graph describing connectivity and influence of each agent over the others. We also consider methods to

compute the opinion change over time by taking into account different cognitive biases of political importance. We considered three kinds of belief updates of significant importance in social networks: The *authority belief* update, which gives more value to the opinion of agents with higher influence, The *confirmation bias* update, which gives more value to the opinion of agents with similar views, and the *back-fire effect* were individuals become more extreme in the opinions when confronted with very different opinions. Based on *concurrent models* from synchronous languages like Esterel and Temporal Concurrent Constraint Programming, we a developed a model for the evolution of polarization in these graphs. In particular, we defined a suitable measure of polarization and characterized graph structures and conditions under which polarization eventually vanishes.

In the work [20] nominated for best paper at CONCUR 2024, we introduced an *asynchronous* DeGroot-based model for opinion dynamics in social networks. The model was formalized using labeled transition systems, henceforth called *opinion transition systems (OTS)*, whose states represented the agents' opinions and whose actions were the edges of the influence graph. If a transition labeled (i,j) is performed, agent j updates their opinion taking into account the opinion of agent i and the influence i had over j. We studied (convergence to) opinion consensus among the agents of strongly-connected graphs with influence values in the interval (0,1). We showed that consensus cannot be guaranteed under the standard strong fairness assumption on transition systems. We derived that consensus is guaranteed under a stronger notion from the literature of concurrent systems: *bounded fairness*. However, we argued that bounded fairness is too strong of a notion for consensus, as it almost surely rules out random runs and was not a constructive liveness property. We introduced a weaker fairness notion, called *m-bounded fairness*, and showed that it guaranteed consensus. The new notion included almost surely all random runs and was a constructive liveness property. Finally, we considered OTS with dynamic influence and showed that convergence to consensus holds under m-bounded fairness if the influence changes within a fixed interval. We illustrated OTS with examples and simulations, offering insights into opinion formation under fairness and dynamic influence.

In [19], we generalized the DeGroot model for opinion dynamics to better capture realistic social scenarios. We introduced a model in which each agent had their own individual cognitive biases. Biases were represented as functions in the square region $[-1,1]$ and categorized into four sub-regions based on the potential reactions they could elicit in an agent during instances of opinion disagreement. Assuming that each agent's bias was a continuous function within the region of receptive but resistant reactions (R), we showed that society converged to a consensus if the graph was strongly connected. Under the same assumption, we established that the entire society converges to a unanimous opinion if and only if the source components of the graph, namely strongly connected components with no external influence, converge to that opinion. We demonstrated that convergence was not guaranteed for strongly connected graphs when biases were either discontinuous functions in R or not included in R. We illustrated our model through a series of examples and simulations, providing insights into how opinions formed in social networks under cognitive biases. More recently in [27] we introduced a generalization of [19] that allows for the modeling of inter-group bias (the tendency to favor the opinion of individual within the same group).

In the work [26] we presented a framework that uses concurrent set relations as the formal basis to specify, simulate, and analyze social interaction systems with dynamic opinion models. Standard models for social learning are obtained as particular instances of the proposed framework. It has been implemented in the Maude system as a fully executable rewrite theory that can be used to better understand how opinions of a system of agents can be shaped. This work also reports an initial exploration in Maude on the use of reachability analysis, probabilistic simulation, and statistical model checking of important properties related to opinion dynamic models.

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

10.1.1 Participation in other International Programs

PROMUEVA

Participants: Frank Valencia, Carlos Pinzon Henao.

Web Page: Project PROMUEVA

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.

Partner Institutions:
- 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

Annabelle McIver

Status Professor

Institution of origin: University of Macquarie

Country: Australia

Dates: January 2024

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

Mobility program/type of mobility: Research stay

Carroll Morgan

Status Professor

Institution of origin: University of New South Wales

Country: Australia

Dates: January 2024

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

Mobility program/type of mobility: Research stay

Fabio Gadducci

Status Professor

Institution of origin: University of Pisa

Country: Italy

Dates: March 2024

Context of the visit: Collaboration with Frank Valencia on polarization.

Mobility program/type of mobility: Research stay

Mario Ferreira Alvim Junior

Status Associate Professor

Institution of origin: Federal University of Mina Gerais (UFMG)

Country: Brazil

Dates: January - March 2024

Context of the visit: Collaboration with Catuscia Palamidessi, Arthur Gaspar Da Silva, Carroll Morgan and Annabelle McIver on privacy and fairness.

Mobility program/type of mobility: Research stay

Judith Sáinz-Pardo Díaz

Status PhD student

Institution of origin: Instituto de Física de Cantabria

Country: Spain

Dates: September - December 2024

Context of the visit: Collaboration with Catuscia Palamidessi, Andreas Athanasiou and Gangsoo Zeong on Privacy in Machine Learning.

Mobility program/type of mobility: Research stay

Arthur Gaspar Da Silva

Status Master student

Institution of origin: Federal University of Mina Gerais (UFMG)

Country: Brazil

Dates: January - March 2024

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

Mobility program/type of mobility: Research stay

10.2.2 Visits to international teams

Frank Valencia

Visited institution: Univ. Javeriana Cali.

Country: Colombia.

Dates: May 01 - May 21, 2024.

Context of the visit: Work on Polarization in Social Networks in the context of the project PROMUEVA.

Mobility program/type of mobility: Research stay.

Frank Valencia

Visited institution: Univ. Javeriana Cali.

Country: Colombia.

Dates: July 01 - August 31, 2024.

Context of the visit: Work on Polarization in Social Networks in the context of the project PROMUEVA.

Mobility program/type of mobility: Research stay.

Frank Valencia

Visited institution: Univ. Javeriana Cali.

Country: Colombia.

Dates: Nov 1 - Nov 21, 2024.

Context of the visit: Work on Polarization in Social Networks 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, Mario Alvim, Sami Zhioua, Ehab ElSalamouni, Héber Arcolezi, Sayan Biswas, Ruta Binkyte-Sadauskiene, Ramon Gonze, Karima Makhlouf.

Web Page: 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 SLU (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 (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 PI: Catuscia Palamidessi

Coordinator: Mario Fritz, CISPA

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, Mario Alvim, Sami Zhioua, Ehab ElSalamouni, Héber Arcolezi, Sayan Biswas, Ruta Binkyte-Sadauskiene, Ramon Gonze, Karima Makhlouf.

Web Page: 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

Participants: Catuscia Palamidessi, Andreas Athanasiou, Konstantino Chatzikokolakis, Gangsoo Zeong.

Web Page: Cryptecs project on cordis.europa.eu

Title: Cloud-Ready Privacy-Preserving Technologies

Duration: From July 1, 2021 to June 30, 2025

Partners:

- Institut National de Recherche en Informatique etc Automatique (Inria), France
- Orange, France
- The Bosch Group, Germany
- University of Stuttgart, Germany
- Zama (SME spin-off of CryptoExperts), France
- Edgeless Systems (SME), Germany

Inria PI: Catuscia Palamidessi

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

Summary: 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

TOBIAS

Participants: Frank Valencia, Carlos Pinzón.

Web Page: TOBIAS

Title: An Interdisciplinary Approach for Testing Opinion Biases in Social Networks

Program: Mission CNRS pour les initiatives transverses et interdisciplinaires (MITI)

Duration: March 2024 - December 2026

Coordinator: Frank Valencia

Partners:

- Cognitive Neuroscience Centre-UMR 5229, Lyon

Inria COMETE contact: Frank Valencia

Description: The project aims to explore the intricate dynamics of opinion formation in social networks by testing and refining our generalization in [19] of the DeGroot model.

iPOP

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

Web Page: Project iPOP

Title: Interdisciplinary Project on Privacy

Program: PEPR Cybersecurity

Duration: 1 October 2022 - 30 September 2028

Coordinator: 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: Project FedMalin

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

Member of organizing committees

- Catuscia Palamidessi has been member of the organizing committee of:
  - APVP 2025, the 15ème Atelier sur la Protection de la Vie Privée (The 15th French Annual Workshop on Privacy). Château du Clos de la Ribaudière, June 9-12, 2025.
  - The ELSA workshop on Privacy Preserving Machine Learning. Bertinoro, Italy, March 16-21, 2025.

11.1.2 Scientific events: selection

Chair of conference program committees

- Catuscia Palamidessi has been chairing the Privacy track of the 32nd ACM Conference on Computer and Communications Security (CCS 2025). Taipei, Taiwan, October 13-17, 2025.

Member of the conference program committees

- Catuscia Palamidessi has been program committee member of:
  - PETS 2025, the International Conference on Privacy Enancing Technologies. Washington DC, USA. July 14–19, 2025.
  - CSF 2025, the International IEEE Symposium on Computer Security Foundations. Santa Cruz, CA, USA. June 16-20, 2025.
  - MFPS 2025 the 41st Conference on Mathematical Foundations of Programming Semantics. Glasgow, Scotland. June 16- 20, 2025,
  - PPAI 2025, the 6th AAAI Workshop on Privacy-Preserving Artificial Intelligence. Pennsylvania Convention Center, Philadelphia, PA, USA. March 3, 2025.

  - NeurIPS 2024, the Thirty-Eighth Annual Conference on Neural Information Processing Systems. Vancouver Convention Center, Canada, December 10-15, 2024.
  - CCS 2024, the 31st ACM Conference on Computer and Communications Security. Salt Lake City, USA. October 14-18, 2024.
  - PETS 2024, the International Conference on Privacy Enancing Technologies. Bristol, UK. July 15–20, 2024.
  - WIL 2024, the Women in Logic workshop. Tallinn, Estonia. July 9, 2024.
  - CSF 2024, the international IEEE Symposium on Computer Security Foundations. Enschede, The Netherlands. July 8-12, 2024.
  - APVP 2024, the 14ème Atelier sur la Protection de la Vie Privée. Domaine Lou Capitelle, France. June 24-27, 2024.
  - FOSSACS 2024, the International Conference on Foundations of Software Science and Computation Structures. Luxembourg City, Luxembourg. April 6–11, 2024.
  - PPAI 2024, the 5th AAAI Workshop on Privacy-Preserving Artificial Intelligence. Vancouver, Canada. February 27, 2024.

- Frank Valencia has been or is a member program committee member of:
  - ICLP-DC 2024. Doctoral Consortium of the 40th International Conference on Logic Programming.
  - EXPRESS/SOS 2024. The 31st International Workshop on Expressiveness in Concurrency
  - COORDINATION 2025. 27th International Conference on Coordination Models and Languages.
  - PPDP 2025. The 27th International Symposium on Principles and Practice of Declarative Programming.

11.1.3 Journal

Member of editorial boards

- Catuscia palamidessi has been member of the editorial board of:
  - (2022-) TheoretiCS.
  - (2020-) Journal of Logical and Algebraic Methods in Programming, Elsevier.
  - (2020-24) IEEE Transactions on Dependable and Secure Computing.
  - (2022-24) ACM Transactions on Privacy and Security.
  - (2015-24) Acta Informatica, Springer.
  - (2006-24) Mathematical Structures in Computer Science, CUP.

11.1.4 Invited talks

- Catuscia Palamidessi has been keynote invited speaker at:
  - WISE 2024, the 25th International Conference on Web Information Systems Engineering. Doha, Qatar. December 2-5, 2024.
  - IT-TML, the ISIT 2024 Workshop on Information-Theoretic Methods for Trustworthy Machine Learning. Athens, Greece. July 7, 2024.
  - APVP 2024, the 14ème Atelier sur la Protection de la Vie Privée. Domaine Lou Capitelle, France. June 24-27, 2024.
  - MFPS XL, the 40th Conference on Mathematical Foundations of Programming Semantics. University of Oxford, UK. June 19-21, 2024.
  - ITASEC 2024, the CINI Conference on Cybersecurity. Salerno, Italy. April 8-11, 2024.

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:
    - (2016-) CONCUR, the International Conference in Concurrency Theory.
    - (2015-) EACSL, the European Association for Computer Science Logics.

11.1.6 Scientific expertise

- Catuscia Palamidessi has been/is:
  - (2025-29) Member of the Scientific Advisory Board of the GSSI international PhD school and a center for research and higher education in Sciences.
  - (2024-25) Member of the international jury of two programs of the FWF, the Austrian Science Fund: the FWF ASTRA Award and the FWF Wittgenstein Award.
  - (2024) Member of the Estonian Research Council for the evaluation process of the research funding applications in 2024, in the fields of Mathematics, Computer Science and Informatics.
  - (2024) Member of the committee for recruitment and promotion of academic staff at Chalmers University of Technology in Sweden.
  - (2024) Member of the Scientific Advisory Board of Digital Futures, a joint lab of the KTH Royal Institute of Technology, the Stockholm University and the RISE Research Institutes of Sweden.
  - (2024) Member of the Scientific Committee of ANR - AAPG 2024. Evaluation of project proposal in the context of Artificial Intelligence and Data Science - CES 23.
  - (2021-) Member of the Board of Trustees of the IMDEA Software Institute, Madrid, Spain.
  - (2019-) Member of the Sci. Adv. Board of CISPA, Helmholtz Center for Information Security. Saarbruecken, Germany.
- Frank Valencia has been:
  - Member of the Selection Committe for the Taltech Postdoctoral Researcher recruitment competiton. November, 2024

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Catuscia Palamidessi has given a tutorial on differential privacy at the Hi! Paris summer school 2024, “AI & Data for Science, Business and Society”, Palaiseau, France, July 2024.
- 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.
- 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

11.2.2 Supervision

Supervision of PhD students

- (2024-) Lois Ecoffet. Co-supervised by Catuscia Palamidessi and by Jean François Couchot, from the Université de Franche-Comteé. Subject: Towards Differentially Private SQL Query Interpretation: A Comprehensive Approach and Implementation in PostgreSQL.
- (2023-) Brahim Erraji. Co-supervised by Catuscia Palamidessi and by Aurélien Bellet, from the Inria team PreMeDICaL. Subject: Fairness in federated learning.
- (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.
- (2021-2024) 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.
- (2023-) Juan Paz. Supervised by Frank Valencia. Subject: Cognitive Bias in Social Networks.

Supervision of postdocs and junior researchers

- (2021-24) Sami Zhioua, researcher CDD.
- (2020-) Gangsoo Zeong, research engineer.
- (2024-) Ehab ElSalamouny, research engineer.
- (2022-) Szilvia Lestyan, postdoc (since 2023 she is hired by the Institut National d'Études Démographiques (INED) and works on a project in the context of a collaboration between INED and COMETE).

11.2.3 Juries

- Catuscia Palamidessi has been:
  – Reviewer and Member of the jury for the HDR defense of Jean Krivine. University of Paris cité, France. January 2025.
  – Reviewer and Member of the jury for the HDR defense of Antoine Boutet. INSA-Lyon, France. December 2024.

- Member of the jury for the HDR defense of Matteo Mio. ENS Lyon, France. April 2024.
- Reviewer and member of the jury for the PhD defense of Dingfan Chen. CISPA, Germany. April 2024.
- Reviewer and member of the jury, in quality of opponent, for the PhD defense of Sara Saeidian. KTH Royal Institute of Technology, Stockholm, Sweden. February 2024.

11.3 Popularization

11.3.1 Productions (articles, videos, podcasts, serious games, ...)

- Frank Valencia co-authored an article on Polarization and Misinformation for the most prestigious Colombian Newspaper *El Tiempo*.

11.3.2 Participation in Live events

- Catuscia Palamidessi has participated as animatrice at the Final of the International Championship of the Mathematical Games organized by the FFJM. Ecole Polytechnique de Paris, France, August 2024.

11.3.3 Others science outreach relevant activities

- Frank Valencia has been interviewed for the following dissemination actions in France and Colombia:

  Magazine: Epsilon Magazine. Réseaux sociaux : polarisent-ils les opinions?. 2024.

  TV: Program Vos Sabes. *Polarización digital*. 2024.

  TV: Telepacífico. *Comprender la polarización para mejorar la toma de decisiones.* 2024.

  TV: Programa En Contacto. *Polarización Política en Redes Sociales*, 2024.

  Website: Agencia de Noticias Univalle. *El Polarizómetro, ciencia e inteligencia artificial para medir la polarización en un mundo influenciado por las redes sociales*. 2024.

  Website: Agencia de Noticias Univalle. *El Polarizómetro, ciencia e inteligencia artificial para medir la polarización en un mundo influenciado por las redes sociales*. 2024.

  Radio: Univalle Radio. *Sanemos Juntos: Ciencia e inteligencia artificial para medir los procesos de polarización en redes sociales.*. 2024.

  Website: Univalle Noticias. *Qué tan polarizados estamos en el Valle del Cauca? Modelo computacional aspira a resolverlo*. 2024.

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. S. Alvim, K. Chatzikokolakis, C. Palamidessi and G. Smith. 'Measuring Information Leakage using Generalized Gain Functions'. In: Computer Security Foundations. Cambridge MA, United States: IEEE, 2012, pp. 265–279. DOI: 10.1109/CSF.2012.26. URL: https://inria.hal.science/hal-00734044.

[3] 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 (cit. on pp. 4, 15).

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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 (cit. on p. 19).

[11] 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 (cit. on p. 7).

12.2 Publications of the year

International journals

[12] H. H. Arcolezi, J.-F. Couchot, B. Al Bouna and X. Xiao. 'Improving the utility of locally differentially private protocols for longitudinal and multidimensional frequency estimates'. In: *Digital Communications and Networks* 10.2 (10th May 2024), pp. 369–379. DOI: 10.1016/j.dcan.2022.07.003. URL: https://inria.hal.science/hal-03727621 (cit. on p. 16).

[13] U. I. Atmaca, S. Biswas, C. Maple and C. Palamidessi. 'A Privacy-Preserving Querying Mechanism with High Utility for Electric Vehicles'. In: *IEEE Open Journal of Vehicular Technology* 5 (30th Jan. 2024), pp. 262–277. DOI: 10.1109/OJVT.2024.3360302. URL: https://hal.science/hal-04467866 (cit. on p. 16).

[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 (2024), pp. 582–596. DOI: 10.56553/popets-2024-0033. URL: https://inria.hal.science/hal-03968692 (cit. on p. 16).

[15] H. H. Arcolezi and S. Gambs. 'Revealing the True Cost of Locally Differentially Private Protocols: An Auditing Perspective'. In: *Proceedings on Privacy Enhancing Technologies* 2024.4 (July 2024), pp. 123–141. DOI: 10.56553/popets-2024-0110. URL: https://inria.hal.science/hal-04644975 (cit. on p. 17).

[16] K. Makhlouf, H. Hwang Arcolezi, S. Zhioua, G. B. Brahim and C. Palamidessi. 'On the Impact of Multi-dimensional Local Differential Privacy on Fairness'. In: *Data Mining and Knowledge Discovery* (27th May 2024), pp. 1–24. DOI: 10.1007/s10618-024-01031-0. URL: https://hal.science/hal-04329938 (cit. on p. 18).

[17] K. Makhlouf, S. Zhioua and C. Palamidessi. 'When causality meets fairness: A survey'. In: *Journal of Logical and Algebraic Methods in Programming* 141 (Oct. 2024), p. 101000. DOI: 10.1016/J.JLAMP.2024.101000. URL: https://inria.hal.science/hal-04950308 (cit. on p. 18).

[18] C. Pinzón, C. Palamidessi, P. Piantanida and F. Valencia. 'On the incompatibility of accuracy and equal opportunity'. In: *Machine Learning* 113.5 (May 2024), pp. 2405–2434. DOI: 10.1007/s10994-023-06331-y. URL: https://hal.science/hal-04308195 (cit. on p. 19).

International peer-reviewed conferences

[19] M. Alvim, A. Gaspar da Silva, S. Knight and F. Valencia. 'A Multi-agent Model for Opinion Evolution in Social Networks Under Cognitive Biases'. In: *Lecture Notes in Computer Science*. FORTE 2024 - 44th International Conference on Formal Techniques for Distributed Objects, Components, and Systems. Vol. 14678. Lecture Notes in Computer Science. Groningen, Netherlands: Springer Nature Switzerland, 13th June 2024, pp. 3–19. DOI: 10.1007/978-3-031-62645-6_1. URL: https://hal.science/hal-04803832 (cit. on pp. 20, 27).

[20] J. Aranda, S. Betancourt, J. Fco and F. Valencia. 'Fairness and Consensus in an Asynchronous Opinion Model for Social Networks'. In: CONCUR 2024 - 35th International Conference on Concurrency Theory. In 35th International Conference on Concurrency Theory (CONCUR 2024). Calgary, Canada: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. DOI: 10.4230/LIPIcs.CONCUR.2024.22. URL: https://hal.science/hal-04803850 (cit. on p. 20).

[21] A. Athanasiou, K. Jung and C. Palamidessi. 'Protection against Source Inference Attacks in Federated Learning using Unary Encoding and Shuffling'. In: *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*. CCS 2024 - The ACM Conference on Computer and Communications Security. Salt Lake City, United States: ACM, 2024, pp. 5036–5038. DOI: 10.1145/3658644.3691411. URL: https://hal.science/hal-04707344 (cit. on p. 17).

[22] R. Binkyte, D. Gorla and C. Palamidessi. 'BaBE: Enhancing Fairness via Estimation of Explaining Variables'. In: FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency. Rio de Janeiro Brazil, France: ACM, 6th May 2024, pp. 1917–1925. DOI: 10.1145/3630106.3659016. URL: https://inria.hal.science/hal-04950351 (cit. on p. 19).

[23] R. Binkyte, C. Pinzón, S. Lestyán, K. Jung, H. Hwang Arcolezi and C. Palamidessi. 'Causal Discovery Under Local Privacy'. In: *Proceedings of Machine Learning Research*. Third Conference on Causal Learning and Reasoning. Vol. 236. Los Angeles, CA, United States, 9th May 2024, pp. 325–383. URL: https://hal.science/hal-04617032 (cit. on pp. 18, 19).

[24] F. Galli, C. Palamidessi and T. Cucinotta. 'Online Sensitivity Optimization in Differentially Private Learning'. In: AAAI Conference on Artificial Intelligence. Vol. 38. Proceedings of the AAAI Conference on Artificial Intelligence 11. Vancouver, Canada, 24th Mar. 2024, pp. 12109–12117. DOI: 10.1609/aaai.v38i11.29099. URL: https://inria.hal.science/hal-04941814 (cit. on p. 17).

[25] K. Makhlouf, T. Stefanović, H. H. Arcolezi and C. Palamidessi. 'A Systematic and Formal Study of the Impact of Local Differential Privacy on Fairness: Preliminary Results'. In: CSF 2024 - 37th IEEE Computer Security Foundations Symposium. Enschede, Netherlands: IEEE, 20th Sept. 2024, pp. 1–16. DOI: 10.1109/CSF61375.2024.00039. URL: https://inria.hal.science/hal-04832154 (cit. on pp. 18, 19).

[26] C. Olarte, C. Ramírez, C. Rocha and F. Valencia. 'Unified Opinion Dynamic Modeling as Concurrent Set Relations in Rewriting Logic'. In: *Lecture Notes in Computer Science*. WRLA 2024 - 15th International Workshop on Rewriting Logic and its Applications. Vol. 14678. Lecture Notes in Computer Science. Luxembourg, Luxembourg: Springer Nature Switzerland, 13th June 2024, pp. 3–19. DOI: 10.1007/978-3-031-65941-6_6. URL: https://hal.science/hal-04803843 (cit. on p. 20).

[27] J. Paz, C. Rocha, L. Tobòn and F. Valencia. 'Consensus in Models for Opinion Dynamics with Generalized-Bias'. In: COMPLEX NETWORKS 2024 - 13th International Conference on Complex Networks & Their Applications. Vol. 12062. Lecture Notes in Computer Science. Istanbul, Turkey: Springer, 2024, pp. 253–269. DOI: 10.48550/arXiv.2409.10809. URL: https://inria.hal.science/hal-04918975 (cit. on p. 20).

Doctoral dissertations and habilitation theses

[28] K. Makhlouf. 'Advancing Ethical and Responsible AI: Exploring Fairness, Privacy, and Explainability through Causal Perspectives'. École polytechnique, 7th Oct. 2024. URL: https://theses.hal.science/tel-04775522.

Reports & preprints

[29] J. Aranda, J. F. Díaz, D. Gaona and F. Valencia. *The Sound of Silence in Social Networks*. 25th Oct. 2024. URL: https://hal.science/hal-04950628.

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

[31] 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.

12.3 Cited publications

[32] 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 (cit. on p. 7).

[33] 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 (cit. on p. 7).

[34] H. H. Arcolezi, K. Makhlouf and C. Palamidessi. '(Local) Differential Privacy has NO Disparate Impact on Fairness'. In: *Lecture Notes in Computer Science*. Vol. LNCS-13942. Proceedings of DBSec 2023 - the 37th IFIP Annual Conference on Data and Applications Security and Privacy. This paper received the Best Paper Award at DBSec 2023. Vijay Atluri and Anna Lisa Ferrara. SOPHIA ANTIPOLIS, France: Springer Nature Switzerland, July 2023, pp. 3–21. DOI: 10.1007/978-3-031-37586-6_1. URL: https://inria.hal.science/hal-04175027 (cit. on p. 19).

[35] 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 (cit. on p. 3).

[36] 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 (cit. on pp. 6, 19).

[37] 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 (cit. on p. 6).

[38] 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 (cit. on p. 8).

[39] 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 (cit. on p. 8).

[40] 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 (cit. on p. 19).

[41] 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 (cit. on p. 6).

[42] 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 (cit. on p. 7).

[43] 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 (cit. on p. 6).

[44] 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 (cit. on p. 6).