RESEARCH CENTRE

Lille - Nord Europe

2021

ACTIVITY REPORT

IN PARTNERSHIP WITH:

CNRS, Université de Lille

Project-Team

MAGNET

Machine Learning in Information Networks

IN COLLABORATION WITH: Centre de Recherche en Informatique, Signal et Automatique de Lille

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Contents

Project-Team MAGNET

Creation of the Project-Team: 2016 May 01

Keywords

Computer sciences and digital sciences

A3.1.3. – Distributed data

A3.1.4. – Uncertain data

A3.4. – Machine learning and statistics

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.4. – Optimization and learning

A3.4.6. – Neural networks

A4.8. – Privacy-enhancing technologies

A9.4. – Natural language processing

Other research topics and application domains

B9.5.1. – Computer science

B9.5.6. – Data science

B9.6.8. – Linguistics

B9.6.10. – Digital humanities

B9.9. – Ethics

B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientists

- Aurelien Bellet [Inria, Researcher]
- Pascal Denis [Inria, Researcher]
- Michael Perrot [Inria, Starting Faculty Position]
- Jan Ramon [Inria, Senior Researcher, HDR]

Faculty Members

- Marc Tommasi [Team leader, Université de Lille, Professor, HDR]
- Mikaela Keller [Université de Lille, Associate Professor]

Post-Doctoral Fellows

- Batiste Le Bars [Inria, from Oct 2021]
- Mohamed Maouche [Inria, until Sep 2021]

PhD Students

- Mahsa Asadi [Inria]
- Moitree Basu [Inria]
- Edwige Cyffers [Université de Lille, from Oct 2021]
- Gaurav Maheshwari [Inria]
- Paul Mangold [Inria]
- Amal Mawass [Université de Lille, from Oct 2021]
- Onkar Pandit [Inria, until Apr 2021]
- Arijus Pleska [Inria]
- César Sabater [Inria]
- Ali Shahin Shamsabadi [Inria, Jan 2021]
- Brij Mohan Lal Srivastava [Inria, until Sep 2021]
- Mariana Vargas Vieyra [Inria, until May 2021]

Technical Staff

- Antoine Barczewski [Inria, Engineer, from Oct 2021]
- Yannick Bouillard [Inria, Engineer, until Jun 2021]
- Marc Damie [Inria, Engineer, from Oct 2021]
- Pradipta Deb [Inria, Engineer, until Feb 2021]
- Joseph Renner [Inria, Engineer]
- Sophie Villerot [Inria, Engineer]

Interns and Apprentices

- Marc Damie [Inria, from Feb 2021 until Jul 2021]
- Marc Andre Sergiel [Université de Lille, from Apr 2021 until Jul 2021]
- Joyce Pascale Tchamdjou Mbotchack [Inria, from Jun 2021 until Aug 2021]
- Maximiliano Vargas Vargas [Inria, from Sep 2021 until Oct 2021]

Administrative Assistants

- Aurore Dalle [Inria, from Apr 2021]
- Julie Jonas [Inria, until Mar 2021]

Visiting Scientist

- Marc Andre Sergiel [Université de Lille, from Sep 2021]

External Collaborator

- Remi Gilleron [Université de Lille, HDR]

2 Overall objectives

The main objective of MAGNET is to develop original machine learning methods for networked data. We consider information networks in which the data consist of feature vectors or texts. We model such networks as graphs wherein nodes correspond to entities (documents, spans of text, users, datasets, learners etc.) and edges correspond to relations between entities (similarity, answer, co-authoring, friendship etc.). In *Mining and Learning in Graphs*, our main research goal is to efficiently search for the best hidden graph structure to be generated for solving a given learning task which exploits the relationships between entities. In *Machine Learning for Natural Language Processing* the objective is to go beyond vectorial classification to solve tasks like coreference resolution and entity linking, temporal structure prediction, and discourse parsing. In *Decentralized Machine Learning* we address the problem of learning in a private, fair and energy efficient way when data are naturally distributed in a network.

The challenges are the dimensionality of the input space, possibly the dimensionality of the output space, the high level of dependencies between the data, the inherent ambiguity of textual data and the limited amount of human labeling. We are interested in making machine learning approaches more acceptable to society. Privacy, sobriety and fairness are important issues that pertain to this research line, and we are interested in the empowerment of end users in the machine learning processes.

3 Research program

The research program of MAGNET is structured along three main axes.

Axis 1: Mining and Learning in Graphs This axis is the backbone of the team. Most of the techniques and algorithms developed in this axis are known by the team members and have impact on the two other axes. We address the following questions and objectives:

How to adaptively build graphs with respect to the given tasks? We study adaptive graph construction along several directions. The first one is to learn the best similarity measure for the graph construction. The second one is to combine different views over the data in the graph construction and learn good representations. We also study weak forms of supervision like comparisons.

How to design methods able to achieve a good trade-off between predictive accuracy and computational complexity? We develop new algorithms for efficient graph-based learning (for instance node prediction or link prediction). In order to deal with scalability issues, our approach is based on optimization, graph sparsification techniques and graph sampling methods.

How to find patterns in graphs based on efficient computations of some statistics? We develop graph mining algorithms and statistics in the context of correlated data.

Axis 2: Machine Learning for Natural Language Processing In this axis, we address the general question that relates graph-based learning and Natural Language Processing (NLP): *How to go beyond vectorial classification models in NLP tasks?* We study the combination of learning representation, structured prediction and graph-based learning methods. Data sobriety and fairness are major constraints we want to deal with. The targeted NLP tasks are coreference resolution and entity linking, temporal structure prediction, and discourse parsing.

Axis 3: Decentralized Machine Learning and Privacy In this axis, we study *How to design private by design machine learning algorithms?* Taking as an opportunity the fact that data collection is now decentralized on smart devices, we propose alternatives to large data centers where data are gathered by developing collaborative and personalized learning.

Contrary to many machine learning approaches where data points and tasks are considered in isolation, we think that a key point of this research is to be able to leverage the relationships between data and learning objectives. Therefore, using graphs as an abstraction of information networks is a major playground for MAGNET. Research related to graph data is a transversal axis, describing a layer of work supporting two other axes on Natural Language Processing and decentralized learning. The machine learning and mining in graphs communities have evolved, for instance taking into account data streams, dynamics but maybe more importantly, focusing on deep learning. Deep neural nets are here to stay, and they are useful tools to tackle difficult problems so we embrace them at different places in the three axes.

MAGNET conducts research along the three axis described above but will put more emphasis on social issues of machine learning. In the context of the recent deployment of artificial intelligence into our daily lives, we are interested in making machine learning approaches more acceptable to society. Privacy, sobriety and fairness are important issues that pertain to this research line, but more generally we are interested in the empowerment of end users in the machine learning processes. Reducing the need of one central authority and pushing more the data processing on the user side, that is decentralization, also participates to this effort. Reducing resources means reducing costs and energy and contributes to building more accessible technologies for companies and users. By considering learning tasks in a more personalized way, but increasing collaboration, we think that we can design solutions that work in low resources regime, with less data or supervision.

In MAGNET we emphasize a different approach than blindly brute-forcing tasks with loads of data. Applications to social sciences for instance have different needs and constraints that motivate data sobriety, fairness and privacy. We are interested in weaker supervision, by leveraging structural properties described in graphs of data, relying on transfer and multi-task learning when faced with graphs of tasks and users. Algorithmic and statistical challenges related to the graph structure of the data still contain open questions. On the statistical side, examples are to take dependencies into account, for instance to compute a mean, to reduce the need of sampling by exploiting known correlations. For the algorithmic point of view, going beyond unlabeled undirected graphs, in particular considering attributed graphs containing text or other information and addressing the case of distributed graphs while maintaining formal guarantees are getting more attention.

In the second axis devoted to NLP, we focus our research on graph-based and representation learning into several directions, all aiming at learning *richer, more robust, and more transferable linguistic representations.* This research program will attempt to bring about strong cross-fertilizations with the other axes, addressing problems in graph, privacy and fairness and making links with decentralized learning. At the intersection between graph-based and representation learning, we will first develop graph embedding algorithms for deriving linguistic representations which are able to capture higher-level semantic and world-knowledge information which eludes strictly distributional models. As an initial step, we envision leveraging pre-existing ontologies (e.g., WordNet, DBpedia), from which one can easily derive interesting similarity graphs between words or noun phrases. We also plan to investigate innovative ways of articulating graph-based semi-supervised learning algorithms and word embedding techniques.

A second direction involves learning representations that are more robust to bias, privacy attacks and adversarial examples. Thus, we intend to leverage recent adversarial training strategies, in which an adversary attempts to recover sensitive attributes (e.g., gender, race) from the learned representations, to be able to neutralize bias or to remove sensitive features. An application domain for this line of research is for instance speech data. The study of learning private representation with its link to fairness in the decentralized setting is another important research topic for the team. In this context of fairness, we also intend to develop similar algorithms for detecting slants, and ultimately for generating de-biased or "re-biased" versions of text embeddings. An illustration is on political slant in written texts (e.g., political speeches and manifestos). Thirdly, we intend to learn linguistic representations that can transfer more easily across languages and domains, in particular in the context of structured prediction problems for low-resource languages. For instance, we first propose to jointly learn model parameters for each language (and/or domains) in a multi-task setting, and leverage a (pre-existing or learned) graph encoding structural similarities between languages (and/or domains). This type of approach would nicely tie in with our previous work on multilingual dependency parsing and on learning personalized models. Furthermore, we will also study how to combine and adapt some neural architectures recently introduced for sequence-to-sequence problems in order to enable transfer of language representations.

In terms of technological transfer, we maintain collaborations with researchers in the humanities and the social sciences, helping them to leverage state-of-the-art NLP techniques to develop new insights to their research by extracting relevant information from large amounts of texts.

The third axis is on distributed and decentralized learning and privacy preserving machine learning. Recent years have seen the evolution of information systems towards ubiquitous computing, smart objects and applications fueled by artificial intelligence. Data are collected on smart devices like smartphones, watches, home devices etc. They include texts, locations, social relationships. Many sensitive data —race, gender, health conditions, tastes etc— can be inferred. Others are just recorded like activities, social relationships but also biometric data like voice and measurements from sensor data. The main tendency is to transfer data into central servers mostly owned by a few tier parties. The situation generates high privacy risks for the users for many reasons: loss of data control, unique entry point for data access, unsolicited data usage etc. But it also increases monopolistic situations and tends to develop oversized infrastructures. The centralized paradigm also has limits when data are too huge such as in the case of multiple videos and sensor data collected for autonomous driving. Partially or fully decentralized systems provide an alternative, to emphasis data exploitation rather than data sharing. For MAGNET, they are source of many new research directions in machine learning at two scales: at the algorithmic level and at a systemic level.

At the algorithmic level the question is to develop new privacy preserving algorithms in the context of decentralized systems. In this context, data remains where it has been collected and learning or statistical queries are processed at the local level. An important question we study is to take into account and measure the impact of collaboration. We also aim at developing methods in the online setting where data arrives continuously or participants join and leave the collaboration network. The granularity of exchanges, the communication cost and the dynamic scenarios, are also studied. On the privacy side, decentralization is not sufficient to establish privacy guarantees because learned models together with the dynamics of collaborative learning may reveal private training data if the models are published or if the communications are observed. But, although it has not been yet well established, decentralization can naturally increase privacy-utility ratio. A direction of research is to formally prove the privacy gain when randomized decentralized protocols are used during learning. In some situations, for instance when part of the data is not sensitive or when trusted servers can be used, a combination between a fully decentralized and a centralized approach is very relevant. In this setting, the question is to find a good trade-off between local versus global computations.

At the systemic layer, in MAGNET we feel that there is a need for research on a global and holistic level, that is to consider full processes involving learning, interacting, predicting, reasoning, repeating etc. rather than studying the privacy of isolated learning algorithms. Our objective is to design languages for describing processes (workflows), data (database schema, background knowledge), population statistics, privacy properties of algorithms, privacy requirements and other relevant information. This is fully aligned with recent trends that aim at giving to statistical learning a more higher level of formal specifications and illustrates our objective for more acceptable and transparent machine learning. We also work towards more robust privacy-friendly systems, being able to handle a wider range of malicious behavior such as collusion to obtain information or inputting incorrect data to obtain information or to influence the result of collaborative computations. From the transfer point of view, we plan to apply transparent, privacy-friendly in significant application domains, such as medicine, surveying, demand prediction and recommendation. In this context, we are interested to understand the appreciation of humans of transparency, verifiability, fairness, privacy-preserving and other trust-increasing aspects of our technologies.

4 Application domains

Our application domain cover health, mobility, social sciences and voice technologies.

Health Privacy is of major importance in the health domain. We contribute to develop methods to give access to the use of data in a private way rather than to the data itself centralized in vulnerable single locations. As an example, we are working with hospitals to develop the means of multicentric studies with privacy guarantees. A second example is personalized medicine where personal devices collect private and highly sensitive data. Potential applications of our research allow to keep data on device and to privately compute statistics.

Social sciences Our NLP research activities are rooted in linguistics, but learning unbiased representations of texts for instance or simply identifying unfair representations also have impacts in political sciences and history.

Voice technologies We develop methods for privacy in speech that can be embedded in software suites dedicated to voice-based interaction systems.

5 Social and environmental responsibility

5.1 Footprint of research activities

Some of our research activities are energy intensive and we will work to reduce this carbon footprint in the future.

5.2 Impact of research results

The main research topics of the team contribute to improve transparency, fairness and privacy in machine learning and reduce bias in natural language processing.

6 Highlights of the year

6.1 Awards

- MICHAËL PERROT was selected as one of the Best Reviewers (Top 10%) at ICML-2021.
- MICHAËL PERROT was selected as one of the Outstanding Reviewers (Top 8%) at NeurIPS-2021.
- Our project on Federated Learning for health with Lille Hospital (team INCLUDE) and 3 other hospitals (G4) has been selected by the CNIL (see the CNIL page).

6.2 Research Activities

- AURÉLIEN BELLET has defended his *habilitation à diriger des recherches*.
- Publication of the well-known and well cited survey on federated learning in *Foundations and Trends in ML* ([5]).

- Magnet will participate in (at least) two projects in the *plan de relance* initiative: Cybersecurity (on privacy); Health (on federated learning).
- PhD defenses: MARIANA VARGAS VIEYRA ([29]), ONKAR PANDIT ([27]), BRIJ MOHAN LAL SRIVASTAVA ([28]).

6.3 Dissemination

- Nijta is a new start-up project by BRIJ MOHAN LAL SRIVASTAVA on voice anonymization.
- A first version of our library on federated learning has been written. It will be used in the context of our partnership with the Lille Hospital.
- PASCAL DENIS was co-chair of TALN-RECITAL 2021, the annual French NLP conference, held in Lille.

7 New software and platforms

7.1 New software

7.1.1 CoRTeX

Name: Python library for noun phrase COreference Resolution in natural language TEXts

Keyword: Natural language processing

Functional Description: CoRTex is a LGPL-licensed Python library for Noun Phrase coreference resolution in natural language texts. This library contains implementations of various state-of-the-art coreference resolution algorithms, including those developed in our research. In addition, it provides a set of APIs and utilities for text pre-processing, reading the CONLL2012 and CONLLU annotation formats, and performing evaluation, notably based on the main evaluation metrics (MUC, B-CUBED, and CEAF). As such, CoRTex provides benchmarks for researchers working on coreference resolution, but it is also of interest for developers who want to integrate a coreference resolution within a larger platform. It currently supports use of the English or French language.

Contact: Pascal Denis

Participant: Pascal Denis

Partner: Orange Labs

7.1.2 Mangoes

Name: MAgnet liNGuistic wOrd vEctorS

Keywords: Word embeddings, NLP

Functional Description: Mangoes is a toolbox for constructing and evaluating static and contextual token vector representations (aka embeddings). The main functionalities are:

- Contextual embeddings: Access a large collection of pretrained transformer-based language models, Pre-train a BERT language model on a corpus, Fine-tune a BERT language model for a number of extrinsic tasks, Extract features/predictions from pretrained language models.

- Static embeddings: Process textual data and compute vocabularies and co-occurrence matrices. Input data should be raw text or annotated text, Compute static word embeddings with different state-of-the art unsupervised methods, Propose statistical and intrinsic evaluation methods, as well as some visualization tools, Generate context dependent embeddings from a pretrained language model.

Future releases will include methods for injecting lexical and semantic knowledge into token and multi-model embeddings, and interfaces into common external knowledge resources.

URL: https://gitlab.inria.fr/magnet/mangoes

Contact: Nathalie Vauquier

7.1.3 metric-learn

Keywords: Machine learning, Python, Metric learning

Functional Description: Distance metrics are widely used in the machine learning literature. Traditionally, practicioners would choose a standard distance metric (Euclidean, City-Block, Cosine, etc.) using a priori knowledge of the domain. Distance metric learning (or simply, metric learning) is the sub-field of machine learning dedicated to automatically constructing optimal distance metrics.

This package contains efficient Python implementations of several popular metric learning algorithms.

URL: https://github.com/scikit-learn-contrib/metric-learn

Contact: Aurelien Bellet

Partner: Parietal

7.1.4 MyLocalInfo

Keywords: Privacy, Machine learning, Statistics

Functional Description: Decentralized algorithms for machine learning and inference tasks which (1) perform as much computation as possible locally and (2) ensure privacy and security by avoiding that personal data leaves devices.

Contact: Nathalie Vauquier

7.1.5 COMPRISE Voice Transformer

Name: COMPRISE Voice Transformer

Keywords: Speech, Privacy

Functional Description: COMPRISE Voice Transformer is an open source tool that increases the privacy of users of voice interfaces by converting their voice into another person's voice without modifying the spoken message. It ensures that any information extracted from the transformed voice can hardly be traced back to the original speaker, as validated through state-of-the-art biometric protocols, and it preserves the phonetic information required for human labelling and training of speech-to-text models.

Release Contributions: This version gives access to the 2 generations of tools that have been used to transform the voice, as part of the COMPRISE project (https://www.compriseh2020.eu/). The first one is a python library that implements 2 basic voice conversion methods, both using VLTN. The second one implements an anonymization method using x-vectors and neural waveform models.

News of the Year: We modified the x-vector based transformer by fixing the percentile-based pitch conversion method, using conda in Docker to fix issues with the Python version, and adding data from the speaker pool to simplify quick start.

URL: https://gitlab.inria.fr/comprise/voice_transformation

Contact: Marc Tommasi

Participants: Nathalie Vauquier, Brij Mohan Lal Srivastava, Marc Tommasi, Emmanuel Vincent, Md Sahidullah

8 New results

8.1 Natural Language Processing

An End-to-End Approach for Full Bridging Resolution, [23]

In this article, we describe our submission to the CODI-CRAC 2021 Shared Task on Anaphora Resolution in Dialogues – Track BR (Gold). We demonstrate the performance of an end-to-end transformer-based higher-order coreference model finetuned for the task of full bridging, that is revealing a relation between mentions. We find that while our approach is not effective at modeling the complexities of the task, it performs well on bridging resolution, suggesting a need for investigations into a robust anaphor identification model for future improvements.

Probing for Bridging Inference in Transformer Language Models, [22]

We probe pre-trained transformer language models for bridging inference. We first investigate individual attention heads in BERT and observe that attention heads at higher layers prominently focus on bridging relations in comparison with the lower and middle layers, also, few specific attention heads concentrate consistently on bridging. More importantly, we consider language models as a whole in our second approach where bridging anaphora resolution is formulated as a masked token prediction task (Of-Cloze test). Our formulation produces optimistic results without any finetuning, which indicates that pre-trained language models substantially capture bridging inference. Our further investigation shows that the distance between anaphor-antecedent and the context provided to language models play an important role in the inference.

What Musical Knowledge Does Self-Attention Learn?, [18, 25]

Since their conception for NLP tasks in 2017, Transformer neural networks have been increasingly used with compelling results for a variety of symbolic Music Information Retrieval (MIR) tasks including music analysis, classification and generation. Although the concept of self-attention between words in text can intuitively be transposed as a relation between musical objects such as notes or chords in a score, it remains relatively unknown what kind of musical relations precisely tend to be captured by self attention mechanisms when applied to musical data. Moreover, the principle of self-attention has been elaborated in NLP to help model the "meaning" of a sentence while in the musical domain this concept appears to be more subjective. In this exploratory work, we open the music transformer black box looking to identify which aspects of music are actually learned by the self-attention mechanism. We apply this approach to two MIR probing tasks : composer classification and cadence identification.

8.2 Decentralized Learning

D-Cliques: Compensating for Data Heterogeneity with Topology in Decentralized Federated Learning, [30]

The convergence speed of machine learning models trained with Federated Learning is significantly affected by heterogeneous data partitions, even more so in a fully decentralized setting without a central server. In this paper, we show that the impact of label distribution skew, an important type of data heterogeneity, can be significantly reduced by carefully designing the underlying communication topology. We present D-Cliques, a novel topology that reduces gradient bias by grouping nodes in sparsely interconnected cliques such that the label distribution in a clique is representative of the global label distribution. We also show how to adapt the updates of decentralized SGD to obtain unbiased gradients and implement an effective momentum with D-Cliques. Our extensive empirical evaluation on MNIST and CIFAR10 demonstrates that our approach provides similar convergence speed as a fully-connected topology, which provides the best convergence in a data heterogeneous setting, with a significant reduction in the number of edges and messages. In a 1000-node topology, D-Cliques require 98% less edges and 96% less total messages, with further possible gains using a small-world topology across cliques.

Federated Multi-Task Learning under a Mixture of Distributions, [20]

The increasing size of data generated by smartphones and IoT devices motivated the development of Federated Learning (FL), a framework for on-device collaborative training of machine learning models. First efforts in FL focused on learning a single global model with good average performance across clients, but the global model may be arbitrarily bad for a given client, due to the inherent heterogeneity of local data distributions. Federated multi-task learning (MTL) approaches can learn personalized models by formulating an opportune penalized optimization problem. The penalization term can capture complex relations among personalized models, but eschews clear statistical assumptions about local data distributions. In this work, we propose to study federated MTL under the flexible assumption that each local data distribution is a mixture of unknown underlying distributions. This assumption encompasses most of the existing personalized FL approaches and leads to federated EM-like algorithms for both client-server and fully decentralized settings. Moreover, it provides a principled way to serve personalized models to clients not seen at training time. The algorithms' convergence is analyzed through a novel federated surrogate optimization framework, which can be of general interest. Experimental results on FL benchmarks show that our approach provides models with higher accuracy and fairness than state-of-the-art methods.

Advances and Open Problems in Federated Learning, [13]

Federated learning (FL) is a machine learning setting where many clients (e.g., mobile devices or whole organizations) collaboratively train a model under the orchestration of a central server (e.g., service provider), while keeping the training data decentralized. FL embodies the principles of focused data collection and minimization, and can mitigate many of the systemic privacy risks and costs resulting from traditional, centralized machine learning and data science approaches. Motivated by the explosive growth in FL research, this monograph discusses recent advances and presents an extensive collection of open problems and challenges.

8.3 Privacy and Machine Learning

Differentially Private Federated Learning on Heterogeneous Data, [35]

Federated Learning (FL) is a paradigm for large-scale distributed learning which faces two key challenges: (i) efficient training from highly heterogeneous user data, and (ii) protecting the privacy of participating users. In this work, we propose a novel FL approach (DP-SCAFFOLD) to tackle these two challenges together by incorporating Differential Privacy (DP) constraints into the popular SCAFFOLD algorithm. We focus on the challenging setting where users communicate with a ”honest-but-curious” server without any trusted intermediary, which requires to ensure privacy not only towards a third-party with access to the final model but also towards the server who observes all user communications. Using advanced results from DP theory, we establish the convergence of our algorithm for convex and non-convex objectives. Our analysis clearly highlights the privacy-utility trade-off under data heterogeneity, and demonstrates the superiority of DP-SCAFFOLD over the state-of-the-art algorithm DP-FedAvg when the number of local updates and the level of heterogeneity grow. Our numerical results confirm our analysis and show that DP-SCAFFOLD provides significant gains in practice.

Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores, [14]

Some organizations such as 23andMe and the UK Biobank have large genomic databases that they re-use for multiple different genome-wide association studies. Even research studies that compile smaller genomic databases often utilize these databases to investigate many related traits. It is common for the study to report a genetic risk score (GRS) model for each trait within the publication. Here, we show that under some circumstances, these GRS models can be used to recover the genetic variants of individuals in these genomic databases-a reconstruction attack. In particular, if two GRS models are trained by using a largely overlapping set of participants, it is often possible to determine the genotype for each of the individuals who were used to train one GRS model, but not the other. We demonstrate this theoretically and experimentally by analyzing the Cornell Dog Genome database. The accuracy of our reconstruction attack depends on how accurately we can estimate the rate of co-occurrence of pairs of single nucleotide polymorphisms within the private database, so if this aggregate information is ever released, it would drastically reduce the security of a private genomic database. Caution should be applied when using the same database for multiple analysis, especially when a small number of individuals are included or excluded from one part of the study.

Mitigating Leakage from Data Dependent Communications in Decentralized Computing using Differential Privacy, [31]

Imagine a group of citizens willing to collectively contribute their personal data for the common good to produce socially useful information, resulting from data analytics or machine learning computations. Sharing raw personal data with a centralized server performing the computation could raise concerns about privacy and a perceived risk of mass surveillance. Instead, citizens may trust each other and their own devices to engage into a decentralized computation to collaboratively produce an aggregate data release to be shared. In the context of secure computing nodes exchanging messages over secure channels at runtime, a key security issue is to protect against external attackers observing the traffic, whose dependence on data may reveal personal information. Existing solutions are designed for the cloud setting, with the goal of hiding all properties of the underlying dataset, and do not address the specific privacy and efficiency challenges that arise in the above context. In this paper, we define a general execution model to control the data-dependence of communications in user-side decentralized computations, in which differential privacy guarantees for communication patterns in global execution plans can be analyzed by combining guarantees obtained on local clusters of nodes. We propose a set of algorithms which allow to trade-off between privacy, utility and efficiency. Our formal privacy guarantees leverage and extend recent results on privacy amplification by shuffling. We illustrate the usefulness of our proposal on two representative examples of decentralized execution plans with data-dependent communications.

Differentially Private Coordinate Descent for Composite Empirical Risk Minimization, [32]

Machine learning models can leak information about the data used to train them. Differentially Private (DP) variants of optimization algorithms like Stochastic Gradient Descent (DP-SGD) have been designed to mitigate this, inducing a trade-off between privacy and utility. In this paper, we propose a new method for composite Differentially Private Empirical Risk Minimization (DP-ERM): Differentially Private proximal Coordinate Descent (DP-CD). We analyze its utility through a novel theoretical analysis of inexact coordinate descent, and highlight some regimes where DP-CD outperforms DP-SGD, thanks to the possibility of using larger step sizes. We also prove new lower bounds for composite DP-ERM under coordinate-wise regularity assumptions, that are, in some settings, nearly matched by our algorithm. In practical implementations, the coordinate-wise nature of DP-CD updates demands special care in choosing the clipping thresholds used to bound individual contributions to the gradients. A natural parameterization of these thresholds emerges from our theory, limiting the addition of unnecessarily large noise without requiring coordinatewise hyperparameter tuning or extra computational cost.

Zero Knowledge Arguments for Verifiable Sampling, [40]

In privacy-preserving machine learning, it is less obvious to verify correct behavior of participants because they are not supposed to reveal their inputs in clear text to other participants. It is hence important to make federated machine learning robust against data poisoning and related attacks. While input data can be related to a distributed ledger (blockchain), a less studied input is formed by the random sampling parties perform. In this paper, we describe strategies based on zero knowledge proofs to allow parties to prove they perform sampling (and other computations) correctly. We sketch a number of alternative ways to implement our idea and provide some preliminary experimental results.

8.4 Federated Learning in Medicine

Specifications for the Routine Implementation of Federated Learning in Hospitals Networks, [19]

We collected user needs to define a process for setting up Federated Learning in a network of hospitals. We identified seven steps: consortium definition, architecture implementation, clinical study definition, data collection, initialization, model training and results sharing. This process adapts certain steps from the classical centralized multicenter framework and brings new opportunities for interaction thanks to the architecture of the Federated Learning algorithms. It is open for completion to cover a variety of scenarios.

8.5 Learning and Speech Recognition

Benchmarking and challenges in security and privacy for voice biometrics, [16]

For many decades, research in speech technologies has focused upon improving reliability. With this now meeting user expectations for a range of diverse applications, speech technology is today omni-present. As result, a focus on security and privacy has now come to the fore. Here, the research effort is in its relative infancy and progress calls for greater, multidisciplinary collaboration with security, privacy, legal and ethical experts among others. Such collaboration is now underway. To help catalyse the efforts, this paper provides a high-level overview of some related research. It targets the non-speech audience and describes the benchmarking methodology that has spearheaded progress in traditional research and which now drives recent security and privacy initiatives related to voice biometrics. We describe: the ASVspoof challenge relating to the development of spoofing countermeasures; the VoicePrivacy initiative which promotes research in anonymisation for privacy preservation.

Enhancing Speech Privacy with Slicing, [33]

Privacy preservation calls for speech anonymization methods which hide the speaker's identity while minimizing the impact on downstream tasks such as automatic speech recognition (ASR) training or decoding. In the recent VoicePrivacy 2020 Challenge, several anonymization methods have been proposed to transform speech utterances in a way that preserves their verbal and prosodic contents while reducing the accuracy of a speaker verification system. In this paper, we propose to further increase the privacy achieved by such methods by segmenting the utterances into shorter slices. We show that our approach has two major impacts on privacy. First, it reduces the accuracy of speaker verification with respect to unsegmented utterances. Second, it also reduces the amount of personal information that can be extracted from the verbal content, in a way that cannot easily be reversed by an attacker. We also show that it is possible to train an ASR system from anonymized speech slices with negligible impact on the word error rate.

Privacy and utility of x-vector based speaker anonymization, [36]

We study the scenario where individuals (speakers) contribute to the publication of an anonymized speech corpus. Data users then leverage this public corpus to perform downstream tasks (such as training automatic speech recognition systems), while attackers may try to de-anonymize it, based on auxiliary knowledge they collect. Motivated by this scenario, speaker anonymization aims to conceal the speaker identity while preserving the quality and usefulness of speech data. In this paper, we study x-vector based speaker anonymization, the leading approach in the recent Voice Privacy Challenge, which converts an input utterance into that of a random pseudo-speaker. We show that the strength of the anonymization varies significantly depending on how the pseudo-speaker is selected. In particular, we investigate four design choices: the distance measure between speakers, the region of x-vector space where the pseudo-speaker is mapped, the gender selection and whether to use speaker or utterance level assignment. We assess the quality of anonymization from the perspective of the three actors involved in our threat model, namely the speaker, the user and the attacker. To measure privacy and utility, we use respectively the linkability score achieved by the attackers and the decoding word error rate incurred by an ASR model trained with the anonymized data. Experiments on LibriSpeech dataset confirm that the optimal combination of design choices yield state-of-the-art performance in terms of privacy protection as well as utility. Experiments on Mozilla Common Voice dataset show that the best design choices with 50 speakers guarantee the same anonymization level against re-identification attack as raw speech with 20,000 speakers.

The VoicePrivacy 2020 Challenge: Results and findings, [39]

This paper presents the results and analyses stemming from the first VoicePrivacy 2020 Challenge which focuses on developing anonymization solutions for speech technology. We provide a systematic overview of the challenge design with an analysis of submitted systems and evaluation results. In particular, we describe the voice anonymization task and datasets used for system development and evaluation. Also, we present different attack models and the associated objective and subjective evaluation metrics. We introduce two anonymization baselines and provide a summary description of the anonymization systems developed by the challenge participants. We report objective and subjective evaluation results for baseline and submitted systems. In addition, we present experimental results for alternative privacy metrics and attack models developed as a part of the post-evaluation analysis. Finally, we summarise our insights and observations that will influence the design of the next VoicePrivacy challenge edition and some directions for future voice anonymization research.

Supplementary material to the paper The VoicePrivacy 2020 Challenge: Results and findings, [38]

The VoicePrivacy 2020 Challenge focuses on developing anonymization solutions for speech technology. This report complements the summary results and analyses presented by Tomashenko et al. (2021). After quickly recalling the challenge design and the submitted anonymization systems, we provide more detailed results and analyses. First, we present objective evaluation results for the primary challenge metrics and for alternative metrics and attack models, and we compare them with each other. Second, we present subjective evaluation results for speaker verifiability, speech naturalness, and speech intelligibility. Finally, we compare these objective and subjective evaluation results with each other.

Study on Acoustic Model Personalization in a Context of Collaborative Learning Constrained by Privacy Preservation, [21]

This paper investigates different approaches in order to improve the performance of a speech recognition system for a given speaker by using no more than 5 min of speech from this speaker, and without exchanging data from other users/speakers. Inspired by the federated learning paradigm, we consider speakers that have access to a personalized database of their own speech, learn an acoustic model and collaborate with other speakers in a network to improve their model. Several local personalizations are explored depending on how aggregation mechanisms are performed. We study the impact of selecting, in an adaptive way, a subset of speakers's models based on a notion of similarity. We also investigate the effect of weighted averaging of fine-tuned and global models. In our approach, only neural acoustic model parameters are exchanged and no audio data is exchanged. By avoiding communicating their personal data, the proposed approach tends to preserve the privacy of speakers. Experiments conducted on the TEDLIUM 3 dataset show that the best improvement is given by averaging a subset of different acoustic models fine-tuned on several user datasets. Our approach applied to HMM/TDNN acoustic models improves quickly and significantly the ASR performance in terms of WER (for instance in one of our two evaluation datasets, from 14.84% to 13.45% with less than 5 min of speech per speaker).

8.6 Fairness and Transparency

For more transparency in the automatic analysis of public consultations: lessons learned from the French "Grand Débat National", [15]

Faced with the limits of representative democracy, digital public consultations provide an opportunity for citizens to contribute their opinions and ideas and for policy makers to involve the population more closely in the public decision making process. The design and deployment of such public consultations pose well-known issues related to potential biases in the questions or in the representativeness of the participants. In this article, we consider the novel issues that arise from the use of artificial intelligence methods to automatically analyze contributions in natural language. Conducting such analyses constitutes a difficult problem for which many approaches (relying on various assumptions and models) exist. Considering the responses to the open-ended questions of the French ”Grand Débat National” as a case study, we show that it is impossible to reproduce the results of the official analysis commissioned by the government. In addition, we identify a number of implicit and arbitrary choices in the official analysis that cast doubts on some of its results. We show also that different methods can lead to different conclusions. Our study highlights the need for greater transparency in the automatic analyses of public consultations so as to ensure reproducibility and public confidence in their results. We conclude with suggestions for improving digital public consultations and their analysis so that they encourage participation and become useful tools for public debate.

Learning Fair Scoring Functions: Bipartite Ranking under ROC-based Fairness Constraints, [24]

Many applications of AI involve scoring individuals using a learned function of their attributes. These predictive risk scores are then used to take decisions based on whether the score exceeds a certain threshold, which may vary depending on the context. The level of delegation granted to such systems in critical applications like credit lending and medical diagnosis will heavily depend on how questions of fairness can be answered. In this paper, we study fairness for the problem of learning scoring functions from binary labeled data, a classic learning task known as bipartite ranking. We argue that the functional nature of the ROC curve, the gold standard measure of ranking accuracy in this context, leads to several ways of formulating fairness constraints. We introduce general families of fairness definitions based on the AUC and on ROC curves, and show that our ROC-based constraints can be instantiated such that classifiers obtained by thresholding the scoring function satisfy classification fairness for a desired range of thresholds. We establish generalization bounds for scoring functions learned under such constraints, design practical learning algorithms and show the relevance our approach with numerical experiments on real and synthetic data.

8.7 Theoretical Computer Science

Linear Programs with Conjunctive Queries, [17]

In this paper, we study the problem of optimizing a linear program whose variables are answers to a conjunctive query. For this we propose the language LP(CQ) for specifying linear programs whose constraints and objective functions depend on the answer sets of conjunctive queries. We contribute an efficient algorithm for solving programs in a fragment of LP(CQ). The naive approach constructs a linear program having as many variables as elements in the answer set of the queries. Our approach constructs a linear program having the same optimal value but fewer variables. This is done by exploiting the structure of the conjunctive queries using hypertree decompositions of small width to group elements of the answer set together. We illustrate the various applications of LP(CQ) programs on three examples: optimizing deliveries of resources, minimizing noise for differential privacy,and computing the s-measure of patterns in graphs as needed for datamining.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

LEGO

Title: LEarning GOod representations for natural language processing

Duration: 2019 ->

Coordinator: Fei Sha (feisha@usc.edu)

Partners:

- University of Southern California

Inria contact: Aurelien Bellet

Summary:

9.1.2 Participation in other International Programs

SLANT: Bilateral ANR project with Luxembourg

Participants: Pascal Denis *(contact person)*, Aurélien Bellet, Mikaela Keller, Gaurav Maheshwari.

Acronym: SLANT

Title: Spin and bias in Language Analyzed in News and Texts

Duration: December 2019 – June 2023

Coordinator: Philippe Muller, IRIT, Toulouse

Partners: IRIT (Toulouse), SnT (Luxembourg)

Abstract: There is a growing concern about misinformation or biased information in public communication, whether in traditional media or social forums. While automating fact-checking has received a lot of attention, the problem of fair information is much larger and includes more insidious forms like biased presentation of events and discussion. The SLANT project aims at characterizing bias in textual data, either intended, in public reporting, or unintended in writing aiming at neutrality. An abstract model of biased interpretation using work on discourse structure, semantics and interpretation will be complemented and concretized by finding relevant lexical, syntactic, stylistic or rhetorical differences through an automated but explainable comparison of texts with different biases on the same subject, based on a dataset of news media coverage from a diverse set of sources. We will also explore how our results can help alter bias in texts or remove it from automated representations of texts.

IMPRESS: Bilateral Inria-DFKI project

Participants: Pascal Denis *(contact person)*, Rémi Gilleron, Priyansh Trivedi.

Acronym: IMPRESS

Title: Improving Embeddings with Semantic Knowledge

Duration: Oct 2020-Sept 2023

Coordinator: Pascal Denis and Ivana Kruijff-Korbayova (DFKI)

Partners: Sémagramme (Inria Nancy), DFKI (Germany)

Abstract: Virtually all NLP systems nowadays use vector representations of words, a.k.a. word embeddings. Similarly, the processing of language combined with vision or other sensory modalities employs multimodal embeddings. While embeddings do embody some form of semantic relatedness, the exact nature of the latter remains unclear. This loss of precise semantic information can affect downstream tasks. Furthermore, while there is a growing body of NLP research on languages other than English, most research on multimodal embeddings is still done on English. The goals of IMPRESS are to investigate the integration of semantic knowledge into embeddings and its impact on selected downstream tasks, to extend this approach to multimodal and mildly multilingual settings, and to develop open source software and lexical resources, focusing on video activity recognition as a practical testbed.

9.2 International research visitors

9.2.1 Visits of international scientists

- Ali Shahin Shamsabadi from Queen Mary University of London has visited Magnet for 6 months. Ali has worked on differential privacy on speech. Research results will be submitted in Jan. 2022.

9.3 European initiatives

9.3.1 FP7 & H2020 projects

COMPRISE

Title: Cost-effective, Multilingual, Privacy-driven voice-enabled Services

Duration: 2018 - 2021

Coordinator: Emmanuel Vincent

Participants: Aurélien Bellet, Marc Tommasi (WP2 Leader), Brij Mohan Lal Srivastava

Partners: ASCORA GMBH (Germany), NETFECTIVE TECHNOLOGY SA (France), ROOTER ANALYSIS SL (Spain), TILDE SIA (Latvia), UNIVERSITAT DES SAARLANDES (Germany), UNIVERSITE DE LORRAINE (France), UNIVERSITE DE LILLE (France)

Inria contact: Akira Campbell

Summary: Besides visual and tactile, the Next Generation Internet will rely more and more on voice interaction. This technology requires huge amounts of speech and language data in every language to reach state-of-the-art performance. The standard today is to store the voices of end users in the cloud and label them manually. This approach raises critical privacy concerns, it limits the number of deployed languages, and it has led to market and data concentration in the hands of big non-European companies such as Google, Facebook, etc.

COMPRISE defines a fully private-by-design methodology and tools that will reduce the cost and increase the inclusiveness of voice interaction technology through research advances on privacy-driven data transformations, personalised learning, automatic labelling, and integrated translation. This leads to a holistic easy-to-use software development kit interoperatingwith a cloud-based resource platform. The sustainability of this new ecosystem will be demonstrated for three sectors with high commercial impact: smart consumer apps, e-commerce, and e-health. COMPRISE will address the mission-oriented challenges of privacy-by-design, inclusiveness, and cost-effectiveness in a sector-agnostic way; allow virtually unlimited collection of real-life non-private quality speech and language data; enable businesses in the Digital Single Market to quickly develop multilingual voice-enabled services in many languages; allow all citizens to transparently access contents and services available in other languages by voice interaction in their own language; result in cost savings for both technology providers and users.

COMPRISE will find application in many sectors beyond those demonstrated, e.g., e-government, e-justice, e-learning, tourism, culture, media, etc. It will have a huge societal impact in terms of unprecedented verifiable privacy guarantees, service to speakers of under-resourced languages or accented speakers, and overall user experience.

9.4 National initiatives

9.4.1 ANR Pamela (2016–2022)

Participants: Marc Tommasi *(contact person)*, Aurélien Bellet, Rémi Gilleron, Jan Ramon, Mahsa Asadi.

The Pamela project aims at developing machine learning theories and algorithms in order to learn local and personalized models from data distributed over networked infrastructures. Our project seeks to provide first answers to modern information systems built by interconnecting many personal devices holding private user data in the search of personalized suggestions and recommendations. More precisely, we will focus on learning in a collaborative way with the help of neighbors in a network. We aim to lay the first blocks of a scientific foundation for these new types of systems, in effect moving from graphs of data to graphs of data and learned models. We argue that this shift is necessary in order to address the new constraints arising from the decentralization of information that is inherent to the emergence of big data. We will in particular focus on the question of learning under communication and privacy constraints. A significant asset of the project is the quality of its industrial partners, Snips and Mediego, who bring in their expertise in privacy protection and distributed computing as well as use cases and datasets. They will contribute to translate this fundamental research effort into concrete outcomes by developing personalized and privacy-aware assistants able to provide contextualized recommendations on small devices and smartphones.

Pamela website

9.4.2 ANR DEEP-Privacy (2019–2023)

Participants: Marc Tommasi *(contact person)*, Aurélien Bellet, Pascal Denis, Jan Ramon, Brij Mohan Lal Srivastava.

DEEP-PRIVACY proposes a new paradigm based on a distributed, personalized, and privacy-preserving approach for speech processing, with a focus on machine learning algorithms for speech recognition. To this end, we propose to rely on a hybrid approach: the device of each user does not share its raw speech data and runs some private computations locally, while some cross-user computations are done by communicating through a server (or a peer-to-peer network). To satisfy privacy requirements at the acoustic level, the information communicated to the server should not expose sensitive speaker information.

9.4.3 ANR-JCJC PRIDE (2020–2025)

Participants: Aurélien Bellet *(contact person)*, Edwige Cyffers, Batiste Le Bars, Paul Mangold.

Machine learning (ML) is ubiquitous in AI-based services and data-oriented scientific fields but raises serious privacy concerns when training on personal data. The starting point of PRIDE is that personal data should belong to the individual who produces it. This requires to revisit ML algorithms to learn from many decentralized personal datasets while preventing the reconstruction of raw data. Differential Privacy (DP) provides a strong notion of protection, but current decentralized ML algorithms are not able to learn useful models under DP. The goal of PRIDE is to develop theoretical and algorithmic tools that enable differentially-private ML methods operating on decentralized datasets, through two complementary objectives: (1) prove that gossip protocols naturally reinforce DP guarantees; (2) propose algorithms at the intersection of decentralized ML and secure multi-party computation.

9.4.4 ANR PMR (2020-2024)

Participant: Jan Ramon *(contact person)*.

Given the growing awareness of privacy risks of data processing, there is an increasing interest in privacy-preserving learning. However, shortcomings in the state of the art limit the applicability of the privacy-preserving learning paradigm. First, most approaches assume too optimistically a honest-but-curious setting. Second, most approaches consider one learning task in isolation, not accounting for the context where querying is a recurring activity. We will investigate new algorithms and models that address these shortcomings. Among others, (i) our algorithms will combine privacy-preserving properties of differential privacy with security offered by cryptography and (ii) based on models of information flows in integrated data handling processes, we will build more refined models analyzing the implications of repeated querying. We will demonstrate the utility of our new theory and algorithms by proposing strategies to realistically apply them in significant real-world problems illustrated through use cases in the medical domain

9.5 Regional initiatives

9.5.1 STaRS (2021-2023)

Participant: Michaël Perrot *(contact person)*.

Machine Learning is becoming ubiquitous in our everyday lives. It is now used in digital assistants, for medical diagnosis, for autonomous vehicles, ... Its success can be explained by the good performances of learned models, sometimes reaching human-level capabilities. However, simply being accurate is not sufficient to ensure that the learning approaches are socially acceptable, in particular if the models are to be largely deployed. Hence, Fairness and Privacy have been extensively studied as standalone trustworthiness notions. However, in practice, it is often mandatory that a model has both properties and thus, jointly studying the two notions is important. This is particularly relevant in decentralized settings where the data is owned by multiple entities that would like to collaborate to learn efficient and fair models but wish to keep their own data private. The goal of this project is twofold: (i) propose new approaches to learn fair and privacy preserving models in a decentralized setting and (ii) provide theoretical guarantees on the trustworthiness level of the learned models that may serve as certificates for the stakeholders.

9.5.2 MusicNLP (2021-)

Participant: Mikaela Keller *(contact person)*.

In the last ten years, deep neural networks have been intensely investigated in the field of Natural Language Processing (NLP). This research has lead to multiple applications including automated corpus annotation and content generation. The temporal nature of music encourages its representation as sequences of elements, most commonly musical notes, that are comparable to sequences of words in NLP. This sequential point of view as well as the common assimilation of music to some kind of language, have motivated the use of NLP approaches for Music Information Retrieval (MIR) tasks, including content analysis and generation. The main goal of this on going project is to evaluate the adaptability, performance and relevance of NLP principles when transposed to the symbolic musical domain. These principles are investigated through the lens of the structural and epistemological differences that exists between natural language and music.

This work is carried out in collaboration with Louis Bigo from CRIStAL Algomus team and was awarded an AIT (Action Incitative Transversale) from CRIStAL.

9.5.3 Federated Learning for multicentric studies

Participants: Aurélien Bellet *(contact person)*, Yannick Bouillard, Paul Mangold, Marc Tommasi.

This regional project is a collaboration between Magnet and Lille Hospital (INCLUDE team) and le Groupement de Coopération Sanitaire G4 (Amiens, Caen, Lille et Rouen hospitals). The aim is to develop and experiment federated learning algorithms for multicentric studies. This project has been awarded by the CNIL.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

- Aurélien Bellet co-organized the Privacy Preserving Machine Learning (PPML'21) workshop at CCS-2021.
- Since May 2020, Aurélien Bellet is co-organizing the Federated Learning One World webinar (900+ registered attendees).

10.1.2 Scientific events: selection

Chair of conference program committees

- Pascal Denis was co-chair of TALN 2021

Member of the conference program committees

- Aurélien Bellet served as Area Chair for ICML-2021, NeurIPS-2021 and AISTATS-2022, and as PC member for FL@ICML-2021, PPAI@AAAI-2022, CAp'21.
- Jan Ramon was PC member of AAAI-2021, AISTATS-2021, DS-2021, ECML-PKDD 2021, ICBINB@NeurIPS 2021, ICDM-2021, ICLR-2021, ICML-2021, IJCAI-2021, ILP-2021, NeurIPS-2021, PriML@NeurIPS-2021, SDM-2021, UAI-2021, XKDD@ECMLPKDD-2021.

- Pascal Denis was PC member of ACL 2021, EMNLP 2021, NAACL 2021. He also part of the inaugural set of Action Editors of the ACL Rolling Review initiative.
- Marc Tommasi was PC member of ICML-2021, NeurIPS-2021, ECML-PKDD-2021 (Area Chair)
- Michaël Perrot served as PC member of ICML-2021, NeurIPS-2021, AISTATS-2022
- Mikaela Keller was PC member of TALN-2021
- Rémi Gilleron served as PC member of ICML-2021, NeurIPS-2021, AISTATS-2022 and ICLR-2022

10.1.3 Journal

Member of the editorial boards

- Jan Ramon is member of the editorial boards of Machine Learning Journal (MLJ), Data Mining and Knowledge Discovery (DMKD), Journal of Machine Learning Research (JMLR), ECML-PKDD Journal track. Jan Ramon is action editor of Data Mining and Knowledge Discovery (DMKD).
- Pascal Denis is standing reviewer for Transactions of the Association for Computational Linguistics (TACL).

Reviewer - reviewing activities

- Jan Ramon also made reviews for journals not mentioned above, e.g., RSIF, KAIS, ICGA
- Aurélien Bellet reviewed for SIAM Journal on Mathematics of Data Science (SIMODS).

10.1.4 Invited talks

- Aurélien Bellet gave invited talks/lectures at Inria Scientific Days 2021, Google Workshop on Federated Learning and Analytics 2021, Google Workshop on Differential Privacy 2021, Federated Learning Workshop 2021, Conférence IA à l'IHP, Hi! Paris Summer School on AI & Data for Science, Business and Society, French-German Summer School on Artificial Intelligence with Industry, Health and Privacy-Preserving Machine Learning Workshop 2021, India-France Knowledge Summit 2021 and seminars at IRT SystemX, GT-PVP, LINC-CNIL, Inria/Inserm HeKA, IMAG Montpellier, Collège industriel de l'AFIA.

10.1.5 Leadership within the scientific community

- Jan Ramon is member of the jury for the 2021 Euroscience young researcher award (EYRA).

10.1.6 Scientific expertise

- Aurélien Bellet was reviewer for scientific proposals at the French National Research Agency (ANR) and for the Hi! Paris Fellowship Program.
- Aurélien Bellet was a member of the recruitment committee of associate professors at Télécom Saint-Etienne and Université Aix-Marseille.
- Jan Ramon participated in proposal review and project monitoring for the H2020 program and the Czech Science Foundation.
- Pascal Denis served as an elected member of the Comité National du CNRS, section 34 (Sciences du Langage).
- Pascal Denis was reviewer for scientific proposals at the French research agency (ANR). Appel A Projet Generique: CE23 - Intelligence Artificielle.
- Marc Tommasi was reviewer for scientific proposals at the French research agency (ANR).

- MIKAELA KELLER was reviewer for scientific proposals at the French research agency (ANR).
- MIKAELA KELLER was a member (scientific expert) of the recruitment committee for an engineer permanent position at INRIA-Lille center.

10.1.7 Research administration

- AURÉLIEN BELLET is member of the Operational Committee for the assesment of Legal and Ethical risks (COERLE).
- MARC TOMMASI is co-head of the DatInG group (4 teams, about 100 persons), member of the Conseil Scientifique du laboratoire CRIStAL and member of the Commission mixte CRIStAL/Faculty of Science, Lille University.
- PASCAL DENIS is a standing member of TALN-RECITAL permanent conference committee ("CPerm").
- PASCAL DENIS is a member of the CNRS GDR NLP Group.
- PASCAL DENIS is administrator of Inria membership to Linguistic Data Consortium (LDC).

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Licence MIASHS: MARC TOMMASI, Data Science, 24h, L2, Université de Lille.
- Licence MIASHS: MARC TOMMASI, Python Programming, 36h, L1 Université de Lille.
- Licence MIASHS: MIKAELA KELLER, Python II, 36h, L2, Université de Lille.
- Licence MIASHS: MIKAELA KELLER, Traitement de données, 24h, L2, Université de Lille.
- Licence SoQ (SHS): MIKAELA KELLER, Algorithmique de graphes, 24h, L3, Université de Lille.
- Master MIASHS: MIKAELA KELLER, Algorithmes fondamentaux de la fouille de données, 27h, M1, Université de Lille.
- Master Data Science: MIKAELA KELLER, Machine Learning 1, 24h, M1, Université de Lille.
- Master Computer Science: MIKAELA KELLER, Apprentissage profond, 24h, M1, Université de Lille.
- Master MIASHS: MICHAËL PERROT, Algorithmes fondamentaux de la fouille de données, 27h, M1, Université de Lille.
- Master Computer Science: MARC TOMMASI, Data Science, 48h, M1, Université de Lille.
- Master Computer Science: MARC TOMMASI, Semi-supervised learning and Graphs, 24h, M2, Université de Lille.
- Master Data Science: MARC TOMMASI Seminars 24h.
- Master Data Science: AURÉLIEN BELLET, Privacy Preserving Machine Learning, 24h, M2, Université de Lille and Ecole Centrale de Lille.
- Master Data Analysis & Decision Making: AURÉLIEN BELLET, Machine Learning, 12h, Ecole Centrale de Lille.
- Master Informatique: PASCAL DENIS, Foundations of Machine Learning, 46h, M1, Université de Lille.
- Master Sciences Cognitives: MICHAËL PERROT, Machine Learning for Cognitive Sciences, 12h, M2, Université de Lille.
- MARC TOMMASI is directeur des études for the Machine Learning master of Computer Science.

10.2.2 Supervision

- Postdoc: Mohamed Maouche. Privacy attacks on representation learning for speech processing, November 2019-October 2021. Aurélien Bellet and Marc Tommasi.
- Postdoc: Batiste Le Bars. On collaboration graph design for decentralized learning. October 2021-. (Aurélien Bellet and Marc Tommasi)
- PhD defended in September 2021: Onkar Pandit, Integrating Contextual and Commonsense Information for Automatic Discourse Understanding. Contributions to Temporal Relation Classification and Bridging Anaphora Resolution, Pascal Denis, Marc Tommasi and Liva Ralaivola (University of Marseille).
- PhD defended in October 2021: Mariana Vargas Vieyra, Graph-based Semi-supervised Learning in Noisy or Missing Graph Settings. Pascal Denis and Aurélien Bellet and Marc Tommasi.
- PhD defended in December 2021: Brij Mohan Lal Srivastava, Representation Learning for Privacy-Preserving Speech Recognition, since Oct 2018 Aurélien Bellet and Marc Tommasi and Emmanuel Vincent.
- PhD in progress: Mahsa Asadi, On Decentralized Machine Learning, since Oct 2018. Aurélien Bellet and Marc Tommasi.
- PhD in progress: Nicolas Crosetti, Privacy Risks of Aggregates in Data Centric-Workflows, since Oct 2018. Florent Capelli and Sophie Tison and Joachim Niehren and Jan Ramon.
- PhD in progress: Moitree Basu, Integrated privacy-preserving AI, since 2019. Jan Ramon.
- PhD in progress: César Sabater, Privacy Preserving Machine Learning, since 2019. Jan Ramon.
- PhD in progress: Paul Mangold. Decentralized Optimization and privacy. Aurélien Bellet and Marc Tommasi and Joseph Salmon, since October 2020.
- Phd in progress: Gaurav Maheshwari. Trustworthy Representations for Natural Language Processing, since Nov 2020. Aurélien Bellet, Mikaela Keller, and Pascal Denis
- Phd in progress: Priyansh Trivedi. Enriching Linguistic Representations with External Knowledge, since Nov 2020. Philippe de Groote (Loria, Nancy) and Pascal Denis
- Phd in progress: Amal Mawass. Privacy preserving statistics and their applications in medicine, since Oct. 2021. Jan Ramon
- Phd in progress: Edwige Cyffers. Decentralized learning and privacy amplification, since Oct. 2021. Aurélien Bellet
- Engineer: Antoine Barczewski. Transparent privacy-preserving machine learning, since Oct 2021. Jan Ramon.
- Engineer: Marc Damie. Secure protocols for verifiable decentralized machine learning, since Oct 2021. Jan Ramon
- Engineer Yannick Bouillard, Federated Learning for multi-centric studies, supervised by Aurélien Bellet Nov. 2020- June 2021.
- Engineer Sophie Villerot, ADT project Tailed: Trustworthy AI Library for Environments which are Decentralized, since Nov. 2020. Jan Ramon
- Engineer Joseph Renner, Improving Word Representations with Semantic Knowledge, since Nov. 2020. Pascal Denis and Rémi Gilleron

10.2.3 Juries

- Marc Tommasi was president of the PhD jury of Laurent Feistauer and the habilitation of Walter Rudametkin
- Marc Tommasi was member of the PhD jury of Raouf Kerkouche, Amaury Bouchra Pilet, Rania Talbi.
- Aurélien Bellet was member of the PhD jury of Raouf Kerkouche, Amaury Bouchra Pilet, Réda Alami, Théo Jourdan and Victor Bouvier.
- Pascal Denis was member of the PhD jury of Rémi Cardon, University of Lille
- Mikaela Keller was member of the PhD jury of Yichang Wang

10.3 Popularization

10.3.1 Internal or external Inria responsibilities

- Jan Ramon was a member of the INRIA-Lille committee emploi-recherche (CER).

10.3.2 Articles and contents

- Aurélien Bellet, Mikaela Keller, Nathalie Vauquier, Pascal Denis, and Rémi Gilleron co-authored an article for The Conversation and Binaire on the use of AI algorithms for analyzing the outcome of online participatory democracy initiatives: see “IA et démocratie participative : comment les réponses au grand débat national ont-elles été analysées ?” and “IA et démocratie participative : la confiance règne ?”.
- Pascal Denis co-authored an article for the CNRS online journal *Lettres InSHS* for its special issue on Computational Linguistics.

11 Scientific production

11.1 Major publications

[1] A. Bellet, R. Guerraoui and H. Hendrikx. ‘Who started this rumor? Quantifying the natural differential privacy guarantees of gossip protocols’. In: *DISC 2020 - 34th International Symposium on Distributed Computing*. Freiburg / Virtual, Germany, Oct. 2020. URL: https://hal.inria.fr/hal-02166432.

[2] A. Bellet, R. Guerraoui, M. Taziki and M. Tommasi. ‘Personalized and Private Peer-to-Peer Machine Learning’. In: *AISTATS 2018 - 21st International Conference on Artificial Intelligence and Statistics*. Lanzarote, Spain, Apr. 2018, pp. 1–20. URL: https://hal.inria.fr/hal-01745796.

[3] M. Dehouck and P. Denis. ‘Delexicalized Word Embeddings for Cross-lingual Dependency Parsing’. In: *EACL*. Vol. 1. EACL 2017. Valencia, Spain, Apr. 2017, pp. 241–250. DOI: 10.18653/v1/E17-1023. URL: https://hal.inria.fr/hal-01590639.

[4] M. Dehouck and P. Denis. ‘Phylogenetic Multi-Lingual Dependency Parsing’. In: *NAACL 2019 - Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, United States, June 2019. URL: https://hal.archives-ouvertes.fr/hal-02143747.

[5] P. Kairouz, B. H. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings et al. ‘Advances and Open Problems in Federated Learning’. In: *Foundations and Trends in Machine Learning* 14.1-2 (2021), pp. 1–210. URL: https://hal.inria.fr/hal-02406503.

[6] O. Kuželka, Y. Wang and J. Ramon. 'Bounds for Learning from Evolutionary-Related Data in the Realizable Case'. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 2016. New York, United States, July 2016. URL: https://hal.archives-ouvertes.fr/hal-01422033.

[7] E. Lassalle and P. Denis. 'Joint Anaphoricity Detection and Coreference Resolution with Constrained Latent Structures'. In: *AAAI Conference on Artificial Intelligence (AAAI 2015)*. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015). Austin, Texas, United States, Jan. 2015. URL: https://hal.inria.fr/hal-01205189.

[8] C. Pelekis, J. Ramon and Y. Wang. 'H\"older-type inequalities and their applications to concentration and correlation bounds'. In: *Indagationes Mathematicae* 28.1 (2017), pp. 170–182. DOI: 10.1016/j.indag.2016.11.017. URL: https://hal.archives-ouvertes.fr/hal-01421953.

[9] T. Ricatte, R. Gilleron and M. Tommasi. 'Skill Rating for Multiplayer Games Introducing Hypernode Graphs and their Spectral Theory'. In: *Journal of Machine Learning Research* 21 (2020), pp. 1–18. URL: https://hal.inria.fr/hal-02566930.

[10] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi and E. Vincent. 'Evaluating Voice Conversion-based Privacy Protection against Informed Attackers'. In: *ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing*. IEEE Signal Processing Society. Barcelona, Spain, May 2020, pp. 2802–2806. URL: https://hal.inria.fr/hal-02355115.

[11] P. Vanhaesebrouck, A. Bellet and M. Tommasi. 'Decentralized Collaborative Learning of Personalized Models over Networks'. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. Fort Lauderdale, Florida., United States, Apr. 2017. URL: https://hal.inria.fr/hal-01533182.

[12] F. Vitale, N. Parotsidis and C. Gentile. 'Online Reciprocal Recommendation with Theoretical Performance Guarantees'. In: *NIPS 2018 - 32nd Conference on Neural Information Processing Systems*. Montreal, Canada, Dec. 2018. URL: https://hal.inria.fr/hal-01916979.

11.2 Publications of the year

International journals

[13] P. Kairouz, B. H. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings et al. 'Advances and Open Problems in Federated Learning'. In: *Foundations and Trends in Machine Learning* 14.1-2 (2021), pp. 1–210. URL: https://hal.inria.fr/hal-02406503.

[14] B. Paige, J. Bell, A. Bellet, A. Gascón and D. Ezer. 'Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores'. In: *Journal of Computational Biology* 28.5 (2021), pp. 435–451. DOI: 10.1089/cmb.2020.0445. URL: https://hal.inria.fr/hal-03498165.

National journals

[15] A. Bellet, P. Denis, R. Gilleron, M. Keller and N. Vauquier. 'For more transparency in the automatic analysis of public consultations: lessons learned from the French "Grand Débat National"'. In: *Statistique et Société*. Gilets jaunes et Grand Débat National : outils, données et analyses 9.1-2 (2021), pp. 147–168. URL: https://hal.inria.fr/hal-02860659.

International peer-reviewed conferences

[16] J.-F. Bonastre, H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, P.-G. Noe, J. Patino, M. Sahidullah, B. M. L. Srivastava, M. Todisco, N. Tomashenko, E. Vincent, X. Wang and J. Yamagishi. 'Benchmarking and challenges in security and privacy for voice biometrics'. In: SPSC 2021 - 1st ISCA Symposium on Security and Privacy in Speech Communication. Magdeburg, Germany, 10th Nov. 2021. DOI: 10.21437/SPSC.2021-11. URL: https://hal.archives-ouvertes.fr/hal-03346196.

[17] F. Capelli, N. Crosetti, J. Niehren and J. Ramon. 'Linear Programs with Conjunctive Queries'. In: 25th Internationcal Conference on Database Theory (ICDT 2022). Edinburgh, United Kingdom, 29th Mar. 2022. URL: https://hal.archives-ouvertes.fr/hal-01981553.

[18] M. Keller, G. Loiseau and L. Bigo. 'What Musical Knowledge Does Self-Attention Learn?' In: Workshop on NLP for Music and Spoken Audio (NLP4MuSA 2021). Online, France, 2021. URL: https://hal.archives-ouvertes.fr/hal-03419236.

[19] A. Lamer, A. Filiot, Y. Bouillard, P. Mangold, P. Andrey and J. Schiro. 'Specifications for the Routine Implementation of Federated Learning in Hospitals Networks'. In: Studies in Health Technology and Informatics, Volume 281: Public Health and Informatics. Virtual Conference, France, 27th May 2021. DOI: 10.3233/shti210134. URL: https://hal.inria.fr/hal-03423328.

[20] O. Marfoq, G. Neglia, A. Bellet, L. Kameni and R. Vidal. 'Federated Multi-Task Learning under a Mixture of Distributions'. In: NeurIPS 2021 - 35th Conference on Neural Information Processing Systems. Sydney, Australia, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03406994.

[21] S. Mdhaffar, M. Tommasi and Y. Estève. 'Study on Acoustic Model Personalization in a Context of Collaborative Learning Constrained by Privacy Preservation'. In: *Speech and Computer 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings.* SPECOM 2021 - 23rd International Conference on Speech and Computer. St Petersburg, Russia, 2021, pp. 426–436. DOI: 10.1007/978-3-030-87802-3_39. URL: https://hal.archives-ouvertes.fr/hal-03369206.

[22] O. Pandit and Y. Hou. 'Probing for Bridging Inference in Transformer Language Models'. In: NAACL 2021 - Annual Conference of the North American Chapter of the Association for Computational Linguistics. Online Conference, Mexico, 6th June 2021. URL: https://hal.inria.fr/hal-03284110.

[23] J. Renner, P. Trivedi, G. Maheshwari, R. Gilleron and P. Denis. 'An End-to-End Approach for Full Bridging Resolution'. In: CODI-CRAC 2021 - Shared-Task: Anaphora Resolution in Dialogues. Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 48–54. URL: https://hal.archives-ouvertes.fr/hal-03431313.

[24] R. Vogel, A. Bellet and S. Clémençon. 'Learning Fair Scoring Functions: Bipartite Ranking under ROC-based Fairness Constraints'. In: AISTATS 2021 - 24th International Conference on Artificial Intelligence and Statistics. Virtual, Unknown Region, 2021. URL: https://hal.inria.fr/hal-03100014.

National peer-reviewed Conferences

[25] M. Keller, K. Akesbi, L. Moreira and L. Bigo. 'Techniques de traitement automatique du langage naturel appliquées aux représentations symboliques musicales'. In: JIM 2021 - Journées d'Informatique Musicale. Virtual, France, 8th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03279850.

Doctoral dissertations and habilitation theses

[26] A. Bellet. 'Contributions to Decentralized and Privacy-Preserving Machine Learning'. Université de Lille, 30th Nov. 2021. URL: https://tel.archives-ouvertes.fr/tel-03542802.

[27] O. Pandit. 'Integrating Contextual and Commonsense Information for Automatic Discourse Understanding :Contributions to Temporal Relation Classification and Bridging Anaphora Resolution'. Université de Lille, 23rd Sept. 2021. URL: https://hal.inria.fr/tel-03528029.

[28] B. M. L. Srivastava. 'Speaker Anonymization: Representation, Evaluation and Formal Guarantees'. Inria Lille Nord Europe - Laboratoire CRIStAL - Université de Lille, 2nd Dec. 2021. URL: https://hal.inria.fr/tel-03539738.

[29] M. Vargas Vieyra. ‘Graph-based Semi-supervised Learning in Missing and Noisy Graph Settings’. Inria Lille Nord Europe - Laboratoire CRIStAL - Université de Lille, 27th Oct. 2021. URL: https://hal.archives-ouvertes.fr/tel-03528596.

Reports & preprints

[30] A. Bellet, A.-M. Kermarrec and E. Lavoie. *D-Cliques: Compensating for Data Heterogeneity with Topology in Decentralized Federated Learning*. 20th Dec. 2021. URL: https://hal.inria.fr/hal-03498160.

[31] R. Ladjel, N. Anciaux, A. Bellet and G. Scerri. *Mitigating Leakage from Data Dependent Communications in Decentralized Computing using Differential Privacy*. 24th Dec. 2021. URL: https://hal.inria.fr/hal-03502320.

[32] P. Mangold, A. Bellet, J. Salmon and M. Tommasi. *Differentially Private Coordinate Descent for Composite Empirical Risk Minimization*. 2nd Feb. 2022. URL: https://hal.inria.fr/hal-03424974.

[33] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi and E. Vincent. *Enhancing Speech Privacy with Slicing*. 7th Oct. 2021. URL: https://hal.inria.fr/hal-03369137.

[34] S. Mdhaffar, J.-F. Bonastre, M. Tommasi, N. Tomashenko and Y. Estève. *RETRIEVING SPEAKER INFORMATION FROM PERSONALIZED ACOUSTIC MODELS FOR SPEECH RECOGNITION*. 22nd Jan. 2022. URL: https://hal.archives-ouvertes.fr/hal-03539741.

[35] M. Noble, A. Bellet and A. Dieuleveut. *Differentially Private Federated Learning on Heterogeneous Data*. 20th Dec. 2021. URL: https://hal.inria.fr/hal-03498158.

[36] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang and J. Yamagishi. *Privacy and utility of x-vector based speaker anonymization*. 28th Dec. 2021. URL: https://hal.inria.fr/hal-03197376.

[37] N. Tomashenko, S. Mdhaffar, M. Tommasi, Y. Estève and J.-F. Bonastre. *Privacy attacks for automatic speech recognition acoustic models in a federated learning framework*. 24th Jan. 2022. URL: https://hal.archives-ouvertes.fr/hal-03539742.

[38] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O’brien, A. Chanclu, J.-F. Bonastre, M. Todisco and M. Maouche. *Supplementary material to the paper The VoicePrivacy 2020 Challenge: Results and findings*. 20th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03335126.

[39] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O’brien, A. Chanclu, J.-F. Bonastre, M. Todisco and M. Maouche. *The VoicePrivacy 2020 Challenge: Results and findings*. 19th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03332224.

Other scientific publications

[40] C. Sabater and J. Ramon. ‘Zero Knowledge Arguments for Verifiable Sampling’. In: NeurIPS 2021 Workshop Privacy in Machine Learning. Sydney (Virtual), Australia, 14th Dec. 2021. URL: https://hal.inria.fr/hal-03464840.

11.3 Other

Scientific popularization

[41] A. Campbell, T. Kleinbauer, M. Tommasi and E. Vincent. ‘Enabling voice-based apps with European values’. In: *ERCIM News* 126 (9th July 2021), pp. 38–39. URL: https://hal.inria.fr/hal-03476390.