IN PARTNERSHIP WITH:
CNRS

Université de Lille

Activity Report 2019

Project-Team MAGNET

Machine Learning in Information Networks

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

RESEARCH CENTER
Lille - Nord Europe

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team MAGNET

Creation of the Team: 2013 January 01, updated into Project-Team: 2016 May 01

Keywords:

Computer Science and Digital Science:

A3.1. - Data
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
A3.5. - Social networks
A3.5.1. - Analysis of large graphs
A3.5.2. - Recommendation systems
A4.8. - Privacy-enhancing technologies
A9.4. - Natural language processing

Other Research Topics and Application Domains:

B1. - Life sciences
B1.1.10. - Systems and synthetic biology
B2. - Health
B2.2.4. - Infectious diseases, Virology
B2.3. - Epidemiology
B2.4.1. - Pharmaco kinetics and dynamics
B2.4.2. - Drug resistance
B5.10. - Biotechnology
B6.3. - Network functions
B7.1.2. - Road traffic
B8.3. - Urbanism and urban planning
B9.5.1. - Computer science
B9.5.4. - Chemistry
B9.5.6. - Data science
B9.6.8. - Linguistics
B9.6.10. - Digital humanities
B9.10. - Privacy

1. Team, Visitors, External Collaborators

Research Scientists

Aurelien Bellet [Inria, Researcher]
Pascal Denis [Inria, Researcher]
Jan Ramon [Inria, Senior Researcher]

Faculty Members

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

Post-Doctoral Fellow

Mohamed Maouche [Inria, from Nov 2019]

PhD Students

Mahsa Asadi [Inria]
Moitree Basu [Inria, from Oct 2019]
Mathieu Dehouck [Université de Lille, from Feb 2019 until Jul 2019]
Onkar Pandit [Inria]
Arijus Pleska [Inria]
César Sabater [Inria]
Brij Mohan Lal Srivastava [Inria]
Mariana Vargas Vieyra [Inria]

Technical staff

William de Vazelhes [Inria, Engineer, until Aug 2019]
Pradipta Deb [Inria, Engineer, from Oct 2019]
Carlos Zubiaga Pena [Inria, Engineer, until Jan 2019]

Intern and Apprentice

Paul Mangold [Ecole Normale Supérieure Lyon, from Oct 2019]

Administrative Assistant

Julie Jonas [Inria]

External Collaborator

Remi Gilleron [Université de Lille, HDR]

2. Overall Objectives

2.1. Presentation

MAGNET is a research group that aims to design new machine learning based methods geared towards mining information networks. Information networks are large collections of interconnected data and documents like citation networks and blog networks among others. Our goal is to propose new prediction methods for texts and networks of texts based on machine learning algorithms in graphs. Such algorithms include node and link classification, link prediction, clustering and probabilistic modeling of graphs. We aim to tackle real-world problems such as browsing, monitoring and recommender systems, and more broadly information extraction in information networks. Application domains cover natural language processing, social networks for cultural data and e-commerce, and biomedical informatics.

3. Research Program

3.1. Introduction

The main objective of MAGNET is to develop original machine learning methods for networked data in order to build applications like browsing, monitoring and recommender systems, and more broadly information extraction in information networks. We consider information networks in which the data consist of both feature vectors and texts. We model such networks as (multiple) (hyper)graphs wherein nodes correspond to entities (documents, spans of text, users, ...) and edges correspond to relations between entities (similarity, answer, co-authoring, friendship, ...). Our main research goal is to propose new online and batch learning algorithms for various problems (node classification / clustering, link classification / prediction) which exploit the relationships between data entities and, overall, the graph topology. We are also interested in searching for the best hidden graph structure to be generated for solving a given learning task. Our research will be based on generative models for graphs, on machine learning for graphs and on machine learning for texts. 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. An additional challenge will be to design scalable methods for large information networks. Hence, we will explore how sampling, randomization and active learning can be leveraged to improve the scalability of the proposed algorithms.

Our research program is organized according to the following questions:

1. How to go beyond vectorial classification models in Natural Language Processing (NLP) tasks?
2. How to adaptively build graphs with respect to the given tasks? How to create networks from observations of information diffusion processes?
3. How to design methods able to achieve a good trade-off between predictive accuracy and computational complexity?
4. How to go beyond strict node homophilic/similarity assumptions in graph-based learning methods?

3.2. Beyond Vectorial Models for NLP

One of our overall research objectives is to derive graph-based machine learning algorithms for natural language and text information extraction tasks. This section discusses the motivations behind the use of graph-based ML approaches for these tasks, the main challenges associated with it, as well as some concrete projects. Some of the challenges go beyond NLP problems and will be further developed in the next sections. An interesting aspect of the project is that we anticipate some important cross-fertilizations between NLP and ML graph-based techniques, with NLP not only benefiting from but also pushing ML graph-based approaches into new directions.

Motivations for resorting to graph-based algorithms for texts are at least threefold. First, online texts are organized in networks. With the advent of the web, and the development of forums, blogs, and micro-blogging, and other forms of social media, text productions have become strongly connected. Interestingly, NLP research has been rather slow in coming to terms with this situation, and most of the literature still focus on document-based or sentence-based predictions (wherein inter-document or inter-sentence structure is not exploited). Furthermore, several multi-document tasks exist in NLP (such as multi-document summarization and cross-document coreference resolution), but most existing work typically ignore document boundaries and simply apply a document-based approach, therefore failing to take advantage of the multi-document dimension [38], [41].

A second motivation comes from the fact that most (if not all) NLP problems can be naturally conceived as graph problems. Thus, NLP tasks often involve discovering a relational structure over a set of text spans (words, phrases, clauses, sentences, etc.). Furthermore, the *input* of numerous NLP tasks is also a graph; indeed, most end-to-end NLP systems are conceived as pipelines wherein the output of one processor is in the input of the next. For instance, several tasks take POS tagged sequences or dependency trees as input. But this structured input is often converted to a vectorial form, which inevitably involves a loss of information.

Finally, graph-based representations and learning methods appear to address some core problems faced by NLP, such as the fact that textual data are typically not independent and identically distributed, they often live on a manifold, they involve very high dimensionality, and their annotations is costly and scarce. As such, graph-based methods represent an interesting alternative to, or at least complement, structured prediction methods (such as CRFs or structured SVMs) commonly used within NLP. Graph-based methods, like label propagation, have also been shown to be very effective in semi-supervised settings, and have already given some positive results on a few NLP tasks [21], [43].

Given the above motivations, our first line of research will be to investigate how one can leverage an underlying network structure (e.g., hyperlinks, user links) between documents, or text spans in general, to enhance prediction performance for several NLP tasks. We think that a “network effect”, similar to the one that took place in Information Retrieval (with the PageRank algorithm), could also positively impact NLP research. A few recent papers have already opened the way, for instance in attempting to exploit Twitter follower graph to improve sentiment classification [42].

Part of the challenge here will be to investigate how adequately and efficiently one can model these problems as instances of more general graph-based problems, such as node clustering/classification or link prediction discussed in the next sections. In a few cases, like text classification or sentiment analysis, graph modeling appears to be straightforward: nodes correspond to texts (and potentially users), and edges are given by relationships like hyperlinks, co-authorship, friendship, or thread membership. Unfortunately, modeling NLP problems as networks is not always that obvious. From the one hand, the right level of representation will probably vary depending on the task at hand: the nodes will be sentences, phrases, words, etc. From the other hand, the underlying graph will typically not be given a priori, which in turn raises the question of how we construct it. A preliminary discussion of the issue of optimal graph construction for semi-supervised learning in NLP is given in [21], [46]. We identify the issue of adaptive graph construction as an important scientific challenge for machine learning on graphs in general, and we will discuss it further in Section 3.3.

As noted above, many NLP tasks have been recast as structured prediction problems, allowing to capture (some of the) output dependencies. How to best combine structured output and graph-based ML approaches is another challenge that we intend to address. We will initially investigate this question within a semi-supervised context, concentrating on graph regularization and graph propagation methods. Within such approaches, labels are typically binary or in a small finite set. Our objective is to explore how one propagates an exponential number of *structured labels* (like a sequence of tags or a dependency tree) through graphs. Recent attempts at blending structured output models with graph-based models are investigated in [43], [31]. Another related question that we will address in this context is how does one learn with *partial labels* (like partially specified tag sequence or tree) and use the graph structure to complete the output structure. This last question is very relevant to NLP problems where human annotations are costly; being able to learn from partial annotations could therefore allow for more targeted annotations and in turn reduced costs [33].

The NLP tasks we will mostly focus on are coreference resolution and entity linking, temporal structure prediction, and discourse parsing. These tasks will be envisioned in both document and cross-document settings, although we expect to exploit inter-document links either way. Choices for these particular tasks is guided by the fact that they are still open problems for the NLP community, they potentially have a high impact for industrial applications (like information retrieval, question answering, etc.), and we already have some expertise on these tasks in the team (see for instance [32], [28], [30]). As a midterm goal, we also plan to work on tasks more directly relating to micro-blogging, such as sentiment analysis and the automatic thread structuring of technical forums; the latter task is in fact an instance of rhetorical structure prediction [45]. We have already initiated some work on the coreference resolution with graph-based learning, by casting the problem as an instance of spectral clustering [30].

3.3. Adaptive Graph Construction

In most applications, edge weights are computed through a complex data modeling process and convey crucially important information for classifying nodes, making it possible to infer information related to each data sample even exploiting the graph topology solely. In fact, a widespread approach to several classification problems is to represent the data through an undirected weighted graph in which edge weights quantify the similarity between data points. This technique for coding input data has been applied to several domains, including classification of genomic data [40], face recognition [29], and text categorization [34].

In some cases, the full adjacency matrix is generated by employing suitable similarity functions chosen through a deep understanding of the problem structure. For example for the TF-IDF representation of documents, the affinity between pairs of samples is often estimated through the cosine measure or the χ^2 distance. After the generation of the full adjacency matrix, the second phase for obtaining the final graph consists in an edge sparsification/reweighting operation. Some of the edges of the clique obtained in the first step are pruned and the remaining ones can be reweighted to meet the specific requirements of the given classification problem. Constructing a graph with these methods obviously entails various kinds of loss of information. However, in problems like node classification, the use of graphs generated from several datasets can lead to an improvement in accuracy ([47], [22], [23]). Hence, the transformation of a dataset into a graph may, at least in some cases, partially remove various kinds of irregularities present in the original datasets, while keeping some of the most useful information for classifying the data samples. Moreover, it is often possible to accomplish classification tasks on the obtained graph using a running time remarkably lower than is needed by algorithms exploiting the initial datasets, and a suitable sparse graph representation can be seen as a compressed version of the original data. This holds even when input data are provided in a online/stream fashion, so that the resulting graph evolves over time.

In this project we will address the problem of adaptive graph construction towards several directions. The first one is about how to choose the best similarity measure given the objective learning task. This question is related to the question of metric and similarity learning ([24], [25]) which has not been considered in the context of graph-based learning. In the context of structured prediction, we will develop approaches where output structures are organized in graphs whose similarity is given by top-k outcomes of greedy algorithms.

A different way we envision adaptive graph construction is in the context of semi-supervised learning. Partial supervision can take various forms and an interesting and original setting is governed by two currently studied applications: detection of brain anomaly from connectome data and polls recommendation in marketing. Indeed, for these two applications, a partial knowledge of the information diffusion process can be observed while the network is unknown or only partially known. An objective is to construct (or complete) the network structure from some local diffusion information. The problem can be formalized as a graph construction problem from partially observed diffusion processes. It has been studied very recently in [36]. In our case, the originality comes either from the existence of different sources of observations or from the large impact of node contents in the network.

We will study how to combine graphs defined by networked data and graphs built from flat data to solve a given task. This is of major importance for information networks because, as said above, we will have to deal with multiple relations between entities (texts, spans of texts, ...) and also use textual data and vectorial data.

3.4. Prediction on Graphs and Scalability

As stated in the previous sections, graphs as complex objects provide a rich representation of data. Often enough the data is only partially available and the graph representation is very helpful in predicting the unobserved elements. We are interested in problems where the complete structure of the graph needs to be recovered and only a fraction of the links is observed. The link prediction problem falls into this category. We are also interested in the recommendation and link classification problems which can be seen as graphs where the structure is complete but some labels on the links (weights or signs) are missing. Finally we are also interested in labeling the nodes of the graph, with class or cluster memberships or with a real value, provided that we have (some information about) the labels for some of the nodes.

The semi-supervised framework will be also considered. A midterm research plan is to study how graph regularization models help for structured prediction problems. This question will be studied in the context of NLP tasks, as noted in Section 3.2, but we also plan to develop original machine learning algorithms that have a more general applicability. Inputs are networks whose nodes (texts) have to be labeled by structures. We assume that structures lie in some manifold and we want to study how labels can propagate in the network. One approach is to find a smooth labeling function corresponding to an harmonic function on both manifolds in input and output.

Scalability is one of the main issues in the design of new prediction algorithms working on networked data. It has gained more and more importance in recent years, because of the growing size of the most popular networked data that are now used by millions of people. In such contexts, learning algorithms whose computational complexity scales quadratically, or slower, in the number of considered data objects (usually nodes or edges, depending on the task) should be considered impractical.

These observations lead to the idea of using graph sparsification techniques in order to work on a part of the original network for getting results that can be easily extended and used for the whole original input. A sparsified version of the original graph can often be seen as a subset of the initial input, i.e. a suitably selected input subgraph which forms the training set (or, more in general, it is included in the training set). This holds even for the active setting. A simple example could be to find a spanning tree of the input graph, possibly using randomization techniques, with properties such that we are allowed to obtain interesting results for the initial graph dataset. We have started to explore this research direction for instance in [44].

At the level of mathematical foundations, the key issue to be addressed in the study of (large-scale) random networks also concerns the segmentation of network data into sets of independent and identically distributed observations. If we identify the data sample with the whole network, as it has been done in previous approaches [35], we typically end up with a set of observations (such as nodes or edges) which are highly interdependent and hence overly violate the classic i.i.d. assumption. In this case, the data scale can be so large and the range of correlations can be so wide, that the cost of taking into account the whole data and their dependencies is typically prohibitive. On the contrary, if we focus instead on a set of subgraphs independently drawn from a (virtually infinite) target network, we come up with a set of independent and identically distributed observations—namely the subgraphs themselves, where subgraph sampling is the underlying ergodic process [26]. Such an approach is one principled direction for giving novel statistical foundations to random network modeling. At the same time, because one shifts the focus from the whole network to a set of subgraphs, complexity issues can be restricted to the number of subgraphs and their size. The latter quantities can be controlled much more easily than the overall network size and dependence relationships, thus allowing to tackle scalability challenges through a radically redesigned approach.

Another way to tackle scalability problems is to exploit the inherent decentralized nature of very large graphs. Indeed, in many situations very large graphs are the abstract view of the digital activities of a very large set of users equipped with their own device. Nowadays, smartphones, tablets and even sensors have storage and computation power and gather a lot of data that serve to analytics, prediction, suggestion and personalized recommendation. Gathering all user data in large data centers is costly because it requires oversized infrastructures with huge energy consumption and large bandwidth networks. Even though cloud architectures can optimize such infrastructures, data concentration is also prone to security leaks, lost of privacy and data governance for end users. The alternative we have started to develop in Magnet is to devise decentralized, private and personalized machine learning algorithms so that they can be deployed in the personal devices. The key challenges are therefore to learn in a collaborative way in a network of learners and to preserve privacy and control on personal data.

3.5. Beyond Homophilic Relationships

In many cases, algorithms for solving node classification problems are driven by the following assumption: linked entities tend to be assigned to the same class. This assumption, in the context of social networks, is known as homophily ([27], [37]) and involves ties of every type, including friendship, work, marriage, age, gender, and so on. In social networks, homophily naturally implies that a set of individuals can be parted into subpopulations that are more cohesive. In fact, the presence of homogeneous groups sharing common interests is a key reason for affinity among interconnected individuals, which suggests that, in spite of its simplicity, this principle turns out to be very powerful for node classification problems in general networks.

Recently, however, researchers have started to consider networked data where connections may also carry a negative meaning. For instance, disapproval or distrust in social networks, negative endorsements on the Web. Although the introduction of signs on graph edges appears like a small change from standard weighted graphs, the resulting mathematical model, called signed graphs, has an unexpectedly rich additional complexity. For example, their spectral properties, which essentially all sophisticated node classification algorithms rely on, are different and less known than those of graphs. Signed graphs naturally lead to a specific inference problem that we have discussed in previous sections: link classification. This is the problem of predicting signs of links in a given graph. In online social networks, this may be viewed as a form of sentiment analysis, since we would like to semantically categorize the relationships between individuals.

Another way to go beyond homophily between entities will be studied using our recent model of hypergraphs with bipartite hyperedges [39]. A bipartite hyperedge connects two ends which are disjoint subsets of nodes. Bipartite hyperedges is a way to relate two collections of (possibly heterogeneous) entities represented by nodes. In the NLP setting, while hyperedges can be used to model bags of words, bipartite hyperedges are associated with relationships between bags of words. But each end of bipartite hyperedges is also a way to represent complex entities, gathering several attribute values (nodes) into hyperedges viewed as records. Our hypergraph notion naturally extends directed and undirected weighted graph. We have defined a spectral theory for this new class of hypergraphs and opened a way to smooth labeling on sets of nodes. The weighting scheme allows to weigh the participation of each node to the relationship modeled by bipartite hyperedges accordingly to an equilibrium condition. This condition provides a competition between nodes in hyperedges and allows interesting modeling properties that go beyond homophily and similarity over nodes (the theoretical analysis of our hypergraphs exhibits tight relationships with signed graphs). Following this competition idea, bipartite hyperedges are like matches between two teams and examples of applications are team creation. The basic tasks we are interested in are hyperedge classification, hyperedge prediction, node weight prediction. Finally, hypergraphs also represent a way to summarize or compress large graphs in which there exists highly connected couples of (large) subsets of nodes.

4. Application Domains

4.1. Application Domains

Our main targeted applications are browsing, monitoring, recommending and mining in information networks. The learning tasks considered in the project such as node clustering, node and link classification and link prediction are likely to yield important improvements in these applications. Application domains cover social networks for cultural data and e-commerce, and biomedical informatics.

We also target applications related to decentralized learning and privacy preserving systems when users or devices are interconnected in large networks. We develop solutions based on urban and mobility data where privacy is a specific requirement.

5. Highlights of the Year

5.1. Highlights of the Year

- Metric-Learn software has been included in the scikit-learn-contrib packages. It records more than 900 stars and 190 forks on GitHub. It is also used by 51 projects.
- AURÉLIEN BELLET has applied for a ERC Starting Grant on privacy-preserving decentralized machine learning.
- MATHIEU DEHOUCK has successfully defended his PhD dissertation on *Multi-Lingual Dependency Parsing: Word Representation and Joint Training for Syntactic Analysis*, and he is doing a post-doc at University of A Coruña (Spain) funded by ERC grant FASTPARSE.
- MARIANA VARGAS VIEYRA's work on probabilistic end-to-end graph-based semi-supervised learning was accepted as one of the 8 contributed talks (among 92 accepted submissions) as the NeurIPS'19 workshop on Graph Representation Learning [1].

6. New Software and Platforms

6.1. CoRTeX

Python library for noun phrase COreference Resolution in natural language TEXts

[1] https://grlearning.github.io/papers/

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.

- Participant: Pascal Denis
- Partner: Orange Labs
- Contact: Pascal Denis
- URL: https://gitlab.inria.fr/magnet/CoRTeX

6.2. Mangoes

MAgnet liNGuistic wOrd vEctorS

KEYWORDS: Word embeddings - NLP

FUNCTIONAL DESCRIPTION: Process textual data and compute vocabularies and co-occurrence matrices. Input data should be raw text or annotated text. Compute word embeddings with different state-of-the art unsupervised methods. Propose statistical and intrinsic evaluation methods, as well as some visualization tools.

- Contact: Nathalie Vauquier
- URL: https://gitlab.inria.fr/magnet/mangoes

6.3. 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.

- Partner: Parietal
- Contact: Aurélien Bellet
- URL: https://github.com/scikit-learn-contrib/metric-learn

6.4. 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 personal data leaves devices.

- Contact: Nathalie Vauquier
- URL: https://gitlab.inria.fr/magnet/mylocalinfo

7. New Results

7.1. Natural Language Processing

Multi-Lingual Dependency Parsing

In [1], MATHIEU DEHOUCK presents his work on Word Representation and Joint Training for Syntactic Analysis. Syntactic analysis is a key step in working with natural languages. With the advances in supervised machine learning, modern parsers have reached human performances. However, despite the intensive efforts of the dependency parsing community, the number of languages for which data have been annotated is still below the hundred, and only a handful of languages have more than ten thousands annotated sentences. In order to alleviate the lack of training data and to make dependency parsing available for more languages, previous research has proposed methods for sharing syntactic information across languages. By transferring models and/or annotations or by jointly learning to parse several languages at once, one can capitalise on languages grammatical similarities in order to improve their parsing capabilities. However, while words are a key source of information for mono-lingual parsers, they are much harder to use in multi-lingual settings because they vary heavily even between very close languages. Morphological features on the contrary, are much more stable across related languages than word forms and they also directly encode syntactic information. Furthermore, it is arguably easier to annotate data with morphological information than with complete dependency structures. With the increasing availability of morphologically annotated data using the same annotation scheme for many languages, it becomes possible to use morphological information to bridge the gap between languages in multi-lingual dependency parsing.

In his thesis, MATHIEU DEHOUCK has proposed several new approaches for sharing information across languages. These approaches have in common that they rely on morphology as the adequate representation level for sharing information. We therefore also introduce a new method to analyse the role of morphology in dependency parsing relying on a new measure of morpho-syntactic complexity. The first method uses morphological information from several languages to learn delexicalised word representations that can then be used as feature and improve mono-lingual parser performances as a kind of distant supervision. The second method uses morphology as a common representation space for sharing information during the joint training of model parameters for many languages. The training process is guided by the evolutionary tree of the various language families in order to share information between languages historically related that might share common grammatical traits. We empirically compare this new training method to independently trained models using data from the Universal Dependencies project and show that it greatly helps languages with few resources but that it is also beneficial for better resourced languages when their family tree is well populated. We eventually investigate the intrinsic worth of morphological information in dependency parsing. Indeed not all languages use morphology as extensively and while some use morphology to mark syntactic relations (via cases and persons) other mostly encode semantic information (such as tense or gender). To this end, we introduce a new measure of morpho-syntactic complexity that measures the syntactic content of morphology in a given corpus as a function of preferential head attachment. We show through experiments that this new measure can tease morpho-syntactic languages and morpho-semantic languages apart and that it is more predictive of parsing results than more traditional morphological complexity measures.

Modal sense classification with task-specific context embeddings Sense disambiguation of modal constructions is a crucial part of natural language understanding. Framed as a supervised learning task, this problem heavily depends on an adequate feature representation of the modal verb context. Inspired by recent work on general word sense disambiguation, we propose in [8] a simple approach of modal sense classification in which standard shallow features are enhanced with task-specific context embedding features. Comprehensive experiments show that these enriched contextual representations fed into a simple SVM model lead to significant classification gains over shallow feature sets.

Learning Rich Event Representations and Interactions for Temporal Relation Classification Most existing systems for identifying temporal relations between events heavily rely on hand-crafted features derived from event words and explicit temporal markers. Besides, less attention has been given to automatically learning con-textualized event representations or to finding complex interactions between events. In [9], we fill this gap in showing that a combination of rich event representations and interaction learning is essential to more accurate temporal relation classification. Specifically, we propose a method in which i) Recurrent Neural Networks (RNN) extract contextual information ii) character embeddings capture morpho-semantic features (e.g. tense, mood, aspect), and iii) a deep Convolutional Neural Network (CNN) finds out intricate interactions between events. We show that the proposed approach outperforms most existing systems on the commonly used dataset while using fully automatic feature extraction and simple local inference.

Phylogenetic Multi-Lingual Dependency Parsing Languages evolve and diverge over time. Their evolutionary history is often depicted in the shape of a phylogenetic tree. Assuming parsing models are representations of their languages grammars, their evolution should follow a structure similar to that of the phylo-genetic tree. In [7], drawing inspiration from multi-task learning, we make use of the phylogenetic tree to guide the learning of multilingual dependency parsers leverag-ing languages structural similarities. Experiments on data from the Universal Dependency project show that phylogenetic training is beneficial to low resourced languages and to well furnished languages families. As a side product of phylogenetic training, our model is able to perform zero-shot parsing of previously unseen languages.

7.2. Decentralized Learning

Trade-offs in Large-Scale Distributed Tuplewise Estimation and Learning The development of cluster computing frameworks has allowed practitioners to scale out various statistical estimation and machine learning algorithms with minimal programming effort. This is especially true for machine learning problems whose objective function is nicely separable across individual data points, such as classification and regression. In contrast, statistical learning tasks involving pairs (or more generally tuples) of data points-such as metric learning, clustering or ranking-do not lend themselves as easily to data-parallelism and in-memory computing. In [13], we investigate how to balance between statistical performance and computational efficiency in such distributed tuplewise statistical problems. We first propose a simple strategy based on occasionally repartitioning data across workers between parallel computation stages, where the number of repartition-ing steps rules the trade-off between accuracy and runtime. We then present some theoretical results highlighting the benefits brought by the proposed method in terms of variance reduction, and extend our results to design distributed stochastic gradient descent algorithms for tuplewise empirical risk minimization. Our results are supported by numerical experiments in pairwise statistical estimation and learning on synthetic and real-world datasets.

Who started this rumor? Quantifying the natural differential privacy guarantees of gossip protocols Gossip protocols, also called rumor spreading or epidemic protocols, are widely used to disseminate information in massive peer-to-peer networks. These protocols are often claimed to guarantee privacy because of the uncertainty they introduce on the node that started the dissemination. But is that claim really true? Can one indeed start a gossip and safely hide in the crowd? In [14], we study gossip protocols using a rigorous mathematical framework based on differential privacy to determine the extent to which the source of a gossip can be traceable. Considering the case of a complete graph in which a subset of the nodes are curious, we derive matching lower and upper bounds on differential privacy showing that some gossip protocols achieve strong privacy guarantees. Our results further reveal an interesting tension between privacy and dissemination speed: the standard “push” gossip protocol has very weak privacy guarantees, while the optimal guarantees are attained at the cost of a drastic increase in the spreading time. Yet, we show that it is possible to leverage the inherent randomness and partial observability of gossip protocols to achieve both fast dissemination speed and near-optimal privacy.

Fully Decentralized Joint Learning of Personalized Models and Collaboration Graphs In [15], we consider the fully decentralized machine learning scenario where many users with personal datasets collaborate to learn models through local peer-to-peer exchanges, without a central coordinator. We propose to train personalized models that leverage a collaboration graph describing the relationships between the users' personal tasks, which we learn jointly with the models. Our fully decentralized optimization procedure alternates between training nonlinear models given the graph in a greedy boosting manner, and updating the collaboration graph (with controlled sparsity) given the models. Throughout the process, users exchange messages only with a small number of peers (their direct neighbors in the graph and a few random users), ensuring that the procedure naturally scales to large numbers of users. We analyze the convergence rate, memory and communication complexity of our approach, and demonstrate its benefits compared to competing techniques on synthetic and real datasets.

Advances and Open Problems in Federated Learning 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, we participated in a collaborative paper [18] that discusses recent advances and presents an extensive collection of open problems and challenges.

7.3. Privacy and Machine Learning

Private Protocols for U-Statistics in the Local Model and Beyond In [16], we study the problem of computing U-statistics of degree 2, i.e., quantities that come in the form of averages over pairs of data points, in the local model of differential privacy (LDP). The class of U-statistics covers many statistical estimates of interest, including Gini mean difference, Kendall's tau coefficient and Area under the ROC Curve (AUC), as well as empirical risk measures for machine learning problems such as ranking, clustering and metric learning. We first introduce an LDP protocol based on quantizing the data into bins and applying randomized response, which guarantees an ϵ-LDP estimate with a Mean Squared Error (MSE) of $O(1/\sqrt{n}\epsilon)$ under regularity assumptions on the U-statistic or the data distribution. We then propose a specialized protocol for AUC based on a novel use of hierarchical histograms that achieves MSE of $O(\alpha^3/n\epsilon^2)$ for arbitrary data distribution. We also show that 2-party secure computation allows to design a protocol with MSE of $O(1/n\epsilon^2)$, without any assumption on the kernel function or data distribution and with total communication linear in the number of users n. Finally, we evaluate the performance of our protocols through experiments on synthetic and real datasets.

Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion? In [11], we study Automatic Speech Recognition (ASR), a key technology in many services and applications. This typically requires user devices to send their speech data to the cloud for ASR decoding. As the speech signal carries a lot of information about the speaker, this raises serious privacy concerns. As a solution, an encoder may reside on each user device which performs local computations to anonymize the representation. In this paper, we focus on the protection of speaker identity and study the extent to which users can be recognized based on the encoded representation of their speech as obtained by a deep encoder-decoder architecture trained for ASR. Through speaker identification and verification experiments on the Librispeech corpus with open and closed sets of speakers, we show that the representations obtained from a standard architecture still carry a lot of information about speaker identity. We then propose to use adversarial training to learn representations that perform well in ASR while hiding speaker identity. Our results demonstrate that adversarial training dramatically reduces the closed-set classification accuracy, but this does not translate into increased open-set verification error hence into increased protection of the speaker identity in practice. We suggest several possible reasons behind this negative result.

Evaluating Voice Conversion-based Privacy Protection against Informed Attackers Speech signals are a rich source of speaker-related information including sensitive attributes like identity or accent. With a small amount of found speech data, such attributes can be extracted and modeled for malicious purposes like voice cloning, spoofing, etc. In [19], we investigate speaker anonymization strategies based on voice conversion. In contrast to prior evaluations, we argue that different types of attackers can be defined depending on the extent of their knowledge about the conversion scheme. We compare two frequency warping-based conversion methods and a deep learning based method in three attack scenarios. The utility of the converted speech is measured through the word error rate achieved by automatic speech recognition, while privacy protection is assessed by state-of-the-art speaker verification techniques (i-vectors and x-vectors). Our results show that voice conversion schemes are unable to effectively protect against an attacker that has extensive knowledge of the type of conversion and how it has been applied, but may provide some protection against less knowledgeable attackers.

7.4. Learning in Graphs

Correlation Clustering with Adaptive Similarity Queries In correlation clustering, we are given n objects together with a binary similarity score between each pair of them. The goal is to partition the objects into clusters so to minimise the disagreements with the scores. In [6], we investigate correlation clustering as an active learning problem: each similarity score can be learned by making a query, and the goal is to minimise both the disagreements and the total number of queries. On the one hand, we describe simple active learning algorithms, which provably achieve an almost optimal trade-off while giving cluster recovery guarantees, and we test them on different datasets. On the other hand, we prove information-theoretical bounds on the number of queries necessary to guarantee a prescribed disagreement bound. These results give a rich characterization of the trade-off between queries and clustering error.

Flattening a Hierarchical Clustering through Active Learning In [12], we investigate active learning by pairwise similarity over the leaves of trees originating from hierarchical clustering procedures. In the realizable setting, we provide a full characterization of the number of queries needed to achieve perfect reconstruction of the tree cut. In the non-realizable setting, we rely on known important-sampling procedures to obtain regret and query complexity bounds. Our algorithms come with theoretical guarantees on the statistical error and, more importantly, lend themselves to linear-time implementations in the relevant parameters of the problem. We discuss such implementations, prove running time guarantees for them, and present preliminary experiments on real-world datasets showing the compelling practical performance of our algorithms as compared to both passive learning and simple active learning baselines.

MaxHedge: Maximising a Maximum Online In [10], we introduce a new online learning framework where, at each trial, the learner is required to select a subset of actions from a given known action set. Each action is associated with an energy value, a reward and a cost. The sum of the energies of the actions selected cannot exceed a given energy budget. The goal is to maximise the cumulative profit, where the profit obtained on a single trial is defined as the difference between the maximum reward among the selected actions and the sum of their costs. Action energy values and the budget are known and fixed. All rewards and costs associated with each action change over time and are revealed at each trial only after the learner's selection of actions. Our framework encompasses several online learning problems where the environment changes over time; and the solution trades-off between minimising the costs and maximising the maximum reward of the selected subset of actions, while being constrained to an action energy budget. The algorithm that we propose is efficient, general and may be specialised to multiple natural online combinatorial problems.

Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast One of the most challenging tasks in modern science is the development of systems biology models: Existing models are often very complex but generally have low predictive performance. The construction of high-fidelity models will require hundreds/thousands of cycles of model improvement, yet few current systems biology research studies complete even a single cycle. In [2], we combined multiple software tools with integrated laboratory robotics to execute three cycles of model improvement of the prototypical eukaryotic cellular transformation, the yeast (Saccharomyces cerevisiae) diauxic shift. In the first cycle, a model outperforming the best previous diauxic shift model was developed using bioinformatic and systems biology tools. In the second cycle, the model was further improved using automatically planned experiments. In the third cycle, hypothesis-led experiments improved the model to a greater extent than achieved using high-throughput experiments. All of the experiments were formalized and communicated to a cloud laboratory automation system (Eve) for automatic execution, and the results stored on the semantic web for reuse. The final model adds a substantial amount of knowledge about the yeast diauxic shift: 92 genes (+45%), and 1 048 interactions (+147%). This knowledge is also relevant to understanding cancer, the immune system, and aging. We conclude that systems biology software tools can be combined and integrated with laboratory robots in closed-loop cycles.

7.5. Metric Learning

Metric learning is at the core of many algorithms for learning graphs. A new software has been published in the scikit-learn contrib repository (See the Software section).

Escaping the Curse of Dimensionality in Similarity Learning: Efficient Frank-Wolfe Algorithm and Generalization Bounds Similarity and metric learning provides a principled approach to construct a task-specific similarity from weakly supervised data. However, these methods are subject to the curse of dimensionality: as the number of features grows large, poor generalization is to be expected and training becomes intractable due to high computational and memory costs. In [3], we propose a similarity learning method that can efficiently deal with high-dimensional sparse data. This is achieved through a parameterization of similarity functions by convex combinations of sparse rank-one matrices, together with the use of a greedy approximate Frank-Wolfe algorithm which provides an efficient way to control the number of active features. We show that the convergence rate of the algorithm, as well as its time and memory complexity, are independent of the data dimension. We further provide a theoretical justification of our modeling choices through an analysis of the generalization error, which depends logarithmically on the sparsity of the solution rather than on the number of features. Our experiments on datasets with up to one million features demonstrate the ability of our approach to generalize well despite the high dimensionality as well as its superiority compared to several competing methods.

metric-learn: Metric Learning Algorithms in Python In [20], we present metric-learn, an open source Python package implementing supervised and weakly-supervised distance metric learning algorithms. As part of scikit-learn-contrib, it provides a unified interface compatible with scikit-learn which allows to easily perform cross-validation, model selection, and pipelining with other machine learning estimators. metric-learn is thoroughly tested and available on PyPi under the MIT licence.

7.6. Graph Algorithms

We collaborate with the Links project team on graph-based computations and evaluation in databases.

Dependency Weighted Aggregation on Factorized Databases In [17], we study a new class of aggregation problems, called dependency weighted aggregation. The underlying idea is to aggregate the answer tuples of a query while accounting for dependencies between them, where two tuples are considered dependent when they have the same value on some attribute. The main problem we are interested in is to compute the dependency weighted count of a conjunctive query. This aggregate can be seen as a form of weighted counting, where the weights of the answer tuples are computed by solving a linear program. This linear program enforces that dependent tuples are not over represented in the final weighted count. The dependency weighted count can be used to compute the s-measure, a measure that is used in data mining to estimate the frequency of a pattern in a graph database. Computing the dependency weighted count of a conjunctive query is NP-hard in general. In this paper, we show that this problem is actually tractable for a large class of structurally restricted conjunctive queries such as acyclic or bounded hypertree width queries. Our algorithm works on a factorized representation of the answer set, in order to avoid enumerating it exhaustively. Our technique produces a succinct representation of the weighting of the answers. It can be used to solve other dependency weighted aggregation tasks, such as computing the (dependency) weighted average of the value of an attribute in the answers set.

7.7. Learning and Speech Recognition

We have worked on privacy and machine learning for speech recognition (See Section 7.3). Additional results concern kernel method for speech recognition.

Kernel Approximation Methods for Speech Recognition In [4], we study the performance of kernel methods on the acoustic modeling task for automatic speech recognition, and compare their performance to deep neural networks (DNNs). To scale the kernel methods to large data sets, we use the random Fourier feature method of Rahimi and Recht (2007). We propose two novel techniques for improving the performance of kernel acoustic models. First, we propose a simple but effective feature selection method which reduces the number of random features required to attain a fixed level of performance. Second, we present a number of metrics which correlate strongly with speech recognition performance when computed on the heldout set; we attain improved performance by using these metrics to decide when to stop training. Additionally, we show that the linear bottleneck method of Sainath et al. (2013a) improves the performance of our kernel models significantly, in addition to speeding up training and making the models more compact. Leveraging these three methods, the kernel methods attain token error rates between 0.5% better and 0.1% worse than fully-connected DNNs across four speech recognition data sets, including the TIMIT and Broadcast News benchmark tasks.

8. Partnerships and Cooperations

8.1. Regional Initiatives

We participate to the *Data Advanced data science and technologies* project (CPER Data). This project is organized following three axes: internet of things, data science, high performance computing. MAGNET is involved in the data science axis to develop machine learning algorithms for big data, structured data and heterogeneous data. The project MyLocalInfo is an open API for privacy-friendly collaborative computing in the internet of things.

MAGNET also has various collaborations with research groups in linguistics and psycholinguistics at Université de Lille, in particular UMR STL (with an ongoing joint ANR project) and UMR SCALab (co-supervision of students).

8.2. National Initiatives

8.2.1. ANR Pamela (2016-2020)

Participants: MARC TOMMASI [correspondent], 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.

https://project.inria.fr/pamela/

8.2.2. ANR JCJC GRASP (2016-2020)

Participants: Pascal Denis [correspondent], Aurélien Bellet, Rémi Gilleron, Mikaela Keller, Marc Tommasi

The GRASP project aims at designing new graph-based Machine Learning algorithms that are better tailored to Natural Language Processing structured output problems. Focusing on semi-supervised learning scenarios, we will extend current graph-based learning approaches along two main directions: (i) the use of structured outputs during inference, and (ii) a graph construction mechanism that is more dependent on the task objective and more closely related to label inference. Combined, these two research strands will provide an important step towards delivering more adaptive (to new domains and languages), more accurate, and ultimately more useful language technologies. We will target semantic and pragmatic tasks such as coreference resolution, temporal chronology prediction, and discourse parsing for which proper Machine Learning solutions are still lacking.

https://project.inria.fr/grasp/

8.2.3. ANR DEEP-Privacy (2019-2023)

Participants: Marc Tommasi [correspondent], Aurélien Bellet, Pascal Denis, Jan Ramon, Brij 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.

8.2.4. ANR-NFS REM (2016-2020)

Participants: Pascal Denis [correspondent], Bo Li, Mathieu Dehouck

With colleagues from the linguistics departments at Université de Lille and University of Neuchâtel (Switzerland), Pascal Denis is a member of another ANR project (REM), funded through the bilateral ANR-NFS Scheme. This project, co-headed by I. Depreatere (Université de Lille) and M. Hilpert (Neufchâtel), proposes to reconsider the analysis of English modal constructions from a multidisciplinary perspective, combining insights from theoretical, psycho-linguistic, and computational approaches.

8.3. European Initiatives

8.3.1. FP7 & H2020 Projects

Participants: Aurelien Bellet, Marc Tommasi, Brij Mohan Lal Srivastava.

Program: H2020 ICT-29-2018 (RIA)

Project acronym: COMPRISE

Project title: Cost-effective, Multilingual, Privacy-driven voice-enabled Services

Duration: Dec 2018 - Nov 2021

Coordinator: Emmanuel Vincent [Inria Nancy - Grand Est]

Other partners: Inria Multispeech, Ascora GmbH, Netfective Technology SA, Rooter Analysis SL, Tilde SIA, University of Saarland

Abstract: COMPRISE will define a fully private-by-design methodology and tools that will reduce the cost and increase the inclusiveness of voice interaction technologies.

8.3.2. Collaborations in European Programs, Except FP7 & H2020

Program: Bilateral ANR project with Luxembourg

Project acronym: SLANT

Project title: Spin and Bias in Language Analyzed in News and Texts

Duration: Dec 2019 - June 2023

Coordinator: Philippe Muller [Université Paul Sabatier]

Other 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.

8.4. International Initiatives

8.4.1. Inria International Labs

Inria@SiliconValley
Associate Team involved in the International Lab:

8.4.1.1. LEGO

Title: LEarning GOod representations for natural language processing

International Partner (Institution - Laboratory - Researcher):

University of Southern California (United States) - Theoretical and Empirical Data Science (TEDS) research group Department of Computer Science - Fei Sha

Start year: 2019

See also: https://team.inria.fr/lego/

LEGO lies in the intersection of Machine Learning and Natural Language Processing (NLP). Its goal is to address the following challenges: what are the right representations for text data and how to learn them in a robust and transferable way? How to apply such representations to solve real-world NLP tasks, specifically in scenarios where linguistic resources are scarce? The past years have seen an increasing interest in learning continuous vectorial embeddings, which can be trained together with the prediction model in an end-to-end fashion, as in recent sequence-to-sequence neural models. However, they are unsuitable to low-resource languages as they require massive amounts of data to train. They are also very prone to overfitting, which makes them very brittle, and sensitive to bias present in the original text as well as to confounding factors such as author attributes. LEGO strongly relies on the complementary expertise of the two partners in areas such as representation learning, structured prediction, graph-based learning, multi-task/transfer learning, and statistical NLP to offer a novel alternative to existing techniques. Specifically, we propose to investigate the following two research directions: (a) optimize the representations to make them robust to bias and adversarial examples, and (b) learn transferable representations across languages and domains, in particular in the context of structured prediction problems for low-resource languages. We will demonstrate the usefulness of the proposed methods on several NLP tasks, including multilingual dependency parsing, machine translation, question answering and text summarization.

8.4.2. Inria Associate Teams Not Involved in an Inria International Labs

North-European Associate Team PAD-ML: Privacy-Aware Distributed Machine Learning.

International Partner: the PPDA team at the Alan Turing Institute.

Start year: 2018

In the context of increasing legislation on data protection (e.g., the recent GDPR), an important challenge is to develop privacy-preserving algorithms to learn from datasets distributed across multiple data owners who do not want to share their data. The goal of this joint team is to devise novel privacy-preserving, distributed machine learning algorithms and to assess their performance and guarantees in both theoretical and practical terms.

8.5. International Research Visitors

8.5.1. Visits of International Scientists

Several international researchers have been invited to give a talk at the MAGNET seminar:

- A. Korba (University College London, UK): Two families of (non-parametric) methods for label ranking
- M. Perrot (Max Planck Institute, Germany): Comparison-Based Learning: Hierarchical Clustering and Classification

8.5.2. Visits to International Teams

8.5.2.1. Research Stays Abroad

- FABIO VITALE was on leave at Department of Computer Science of Sapienza University (Rome, Italy) in the Algorithms Randomization Computation group with Prof. Alessandro Panconesi and Prof. Flavio Chierichetti. His current work on machine learning in graphs and published the following papers [6], [12], [10].
- AURÉLIEN BELLET and CÉSAR SABATER visited the Alan Turing Institute (London, UK) for one week in March 2019. They worked with Adrià Gascón, Brooks Paige, Daphne Ezer and Matt Kusner on privacy-preserving machine learning and privacy attacks in genomics.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events: Organisation

9.1.1.1. Member of the Organizing Committees

- Marc Tommasi and Aurélien Bellet co-organized the APVP conference on privacy [2].

9.1.2. Scientific Events: Selection

9.1.2.1. Member of the Conference Program Committees

- Marc Tommasi served as PC member for ICML'19, CAP'19, IJCAI'19 (Senior PC chair), ECML'19 (Area Chair), NeurIPS'19 (Area Chair), APVP'19.
- Jan Ramon served as PC member for AAAI'19, AISTATS'19 & '20, APVP'19, Bigdata'19, CIKM'19, DS'19, ECML/PKDD'19, GEM@ECML'19, IEEE-ICDM'19, ICML'19, IJCAI'19, ILP'19, LEG@ECML'19, LOD'19, MLG'19, NeurIPS'19, SDM'19 & '20.
- Pascal Denis served as PC member for ACL'19 and EMNLP'19.
- Aurélien Bellet served as Area Chair for ICML'19, and PC member for AISTATS'20, NeurIPS'19, PPML workshop at CSS'19, PPML workshop at NeurIPS'19, FL workshop at NeurIPS'19, PPAI workshop at AAAI '20, CAP'19.
- Mikaela Keller served as PC member for ICML'19, CAP'19 and ECML'19.
- Rémi Gilleron served as PC member for AISTATS'19 & '20, ICLR'19 & '20, ECML/PKDD'19, NeurIPS'19 and CAP'19.

9.1.3. Journal

9.1.3.1. Member of the Editorial Boards

- Jan Ramon was member of the editorial board of Data mining and knowledge discovery (DMKD), and of the editorial board of Machine learning journal (MLJ)

9.1.3.2. Reviewer - Reviewing Activities

- Marc Tommasi was reviewer for the ECML-PKDD Journal Track.
- Aurélien Bellet was reviewer for SIAM Journal on Mathematics of Data Science.

9.1.4. Invited Talks

- Aurélien Bellet gave invited talks at several workshops and research school: "Graph signals: learning and optimization perspectives" [3], "The power of graphs in machine learning and sequential decision making" [4], "CIFAR-UKRI-CNRS Workshop on AI & Society: From principles to practice" [5], "Workshop on Federated Learning and Analytics" [6], "International Workshop on Machine Learning & Artificial Intelligence" [7], "Kick-off seminar of the HumAIn Alliance in Artificial Intelligence" [8], "Research School on Uncertainty in Scientific Computing" [9], "GDR RSD and ASF Winter School on Distributed Systems and Networks" [10].
- Aurélien Bellet was invited to talk at the seminars of Thales Research & Technology, Miles team at Paris Dauphine.
- Mikaela Keller gave an invited talk on "Algorithmiques de graphes pour l'étude de réseaux sociaux" at "École d'été internationale : Formation de maîtrise et de doctorat en méthodes et outils numériques : l'analyse de réseaux en histoire" [11].

[2] https://project.inria.fr/apvp2019/

[3] https://graph-sig-2019.sciencesconf.org/

[4] https://graphpower.inria.fr/

[5] https://www.turing.ac.uk/events/cifar-ukri-cnrs-ai-society-principles-practice

[6] https://sites.google.com/view/federated-learning-2019/

[7] https://workshopmlai.wp.imt.fr/

[8] https://www.alliance-humain.fr/

[9] http://www.gdr-mascotnum.fr/etics.html

[10] https://sites.google.com/site/rsdwinterschool/

[11] http://www.grhs.uqam.ca/?portfolio=ecole-dete-internationale-formation-de-maitrise-et-de-doctorat-methodes-et-outils-numeriques-lanalyse-de-reseaux-en-histoire

9.1.5. Scientific Expertise

- Aurélien Bellet was a member of the jury for the Gilles-Kahn PhD award of the French Society of Computer Science (SIF), sponsored by the French Academy of Sciences [12].
- Jan Ramon was a member of the jury for the European Young Researchers Award (EYRA) of Euroscience.
- Pascal Denis served as CoCNRS representative for the evaluation panel of Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur, HCERES [13].
- Pascal Denis acted as external expert for Innovis [14], Brussels Institute for Research and Innovation (Belgium).

9.1.6. Research Administration

- Mikaela Keller is member of the Conseil du laboratoire CRIStAL.
- Marc Tommasi is co-head of the DatInG group (4 teams, 80 persons) and member of the Conseil Scientifique du laboratoire CRIStAL.
- Pascal Denis served as an elected member of the Comité National du CNRS, section 34 (Sciences du Langage).
- Pascal Denis served as a member of the CNRS Pre-GDR NLP Group.
- Jan Ramon was a member of the Inria-Lille committee emploi-recherche (CER).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence SCE: Fabio Vitale, Apprentissage et émergence des comportements, 30h, L2, Université de Lille.

Licence MIASHS: Fabio Vitale, Algorithmique et programmation, 42h, L3, Université de Lille.

Licence SoQ: Fabio Vitale, Bases de données et SQL, 16h, L1, Université de Lille.

Licence MIASHS: Fabio Vitale, Introduction aux bases de données, 18h, L3, Université de Lille.

Licence SHS: Marc Tommasi, Data Science, 24h, L2, Université de Lille.

Licence MIASHS: Marc Tommasi, Codage et représentation de l'information, 48h, L1, Université de Lille.

Licence: Marc Tommasi, Traitement de textes et tableur, 28h Université de Lille.

Licence: Marc Tommasi, Humanités numériques - Découvrir et faire découvrir la programmation, 20h, Université de Lille

Master SHS: Marc Tommasi, IT Tools for professional translators, 24h, M2 Université de Lille.

Licence SHS: Marc Tommasi, Python Programming, 48h, L2 Université de Lille.

Licence MIASHS: Mikaela Keller, Python II, 40h, L2, Université de Lille.

Licence MIASHS: Mikaela Keller, Traitement de données, 42h, 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, 60h, M1, Université de Lille.

Master Computer Science: Mikaela Keller, Machine Learning, 24h, M1, Université de Lille.

Master Data Science: Fabio Vitale, Algorithms and their complexity, 30h, M1, Ecole Centrale de Lille.

[12] https://www.societe-informatique-de-france.fr/recherche/prix-de-these-gilles-kahn/

[13] https://www.hceres.fr/en

[14] https://innoviris.brussels/

Master Data Analysis & Decision Making: AURÉLIEN BELLET, Machine Learning, 12h, Ecole Centrale de Lille.

Master / Master Spécialisé Big Data: AURÉLIEN BELLET, Advanced Machine Learning, 15h, Télécom ParisTech.

Formation continue (Certificat d'Études Spécialisées Data Scientist): AURÉLIEN BELLET, Supervised Learning and Support Vector Machines, 14h, Télécom ParisTech.

Master Informatique: PASCAL DENIS, Foundations of Machine Learning, 46h, M1, Université de Lille.

9.2.2. Supervision

Postdoc: MOHAMED MAOUCHE, supervised by AURÉLIEN BELLET, MARC TOMMASI, Privacy attacks on representation learning for speech processing, since November 2019.

PhD: MATHIEU DEHOUCK, Multi-lingual Dependency Parsing: Word Representation and Joint Training for Syntactic Analysis, 2015-2019, PASCAL DENIS and MARC TOMMASI.

PhD in progress: ONKAR PANDIT, Graph-based Semi-supervised Linguistic Structure Prediction, since Dec. 2017, PASCAL DENIS, MARC TOMMASI and LIVA RALAIVOLA (University of Marseille).

PhD in progress: MARIANA VARGAS VIEYRA, Adaptive Graph Learning with Applications to Natural Language Processing, since Jan. 2018. PASCAL DENIS and AURÉLIEN BELLET and MARC TOMMASI.

PhD in progress: BRIJ 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: ROBIN VOGEL, Learning to rank by similarity and performance optimization in biometric identification, since 2017 (CIFRE thesis with IDEMIA and Télécom ParisTech). AURÉLIEN BELLET, STÉPHAN CLÉMENÇON and ANNE SABOURIN.

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.

Master internship in progress: PAUL MANGOLD, supervised by AURÉLIEN BELLET and MARC TOMMASI, since October 2019.

9.2.3. Juries

MARC TOMMASI was member of the PhD jury of Corentin Hardy (rapporteur), Guillaume Metzler (rapporteur), Ronan Fruit (President)

MARC TOMMASI was member of the Habilitation jury of Rémi Eyraud (rapporteur).

MARC TOMMASI was member of the recruitment committee of Assistant Professors in Computer Science at École centrale de Lille.

AURÉLIEN BELLET was member of the PhD jury of Alexandre Garcia (Télécom Paris).

MIKAELA KELLER was member of the PhD jury of Guillaume Lample (LIP6)

MIKAELA KELLER was member of the recruitment committee of Assistant Professors in Computer Science at Sorbonne Université (LIP6).

RÉMI GILLERON was head of the PhD jury of Lily Galois (Université de Lille)

9.3. Popularization

9.3.1. Internal or external Inria responsibilities

Aurélien Bellet is member of the Operational Committee for the assesment of Legal and Ethical risks (COERLE).

Pascal Denis is administrator of Inria membership to Linguistic Data Consortium (LDC).

9.3.2. Articles and contents

Aurélien Bellet and Marc Tommasi provided expertise for the Dopamine TV program on Arte about new technologies [15].

Aurélien Bellet co-authored an article in the French daily newspaper Libération on AI and transparency [16].

Rémi Gilleron authored a book (in French) introducing Machine Learning [17].

Rémi Gilleron gave a talk in order to popularize Artificial Intelligence.

9.3.3. Interventions

In educational institutions: Marc Tommasi was facilitator for the event *Jouer à débattre*, about machine learning in high school (Tourcoing).

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] M. Dehouck. *Multi-Lingual Dependency Parsing : Word Representation and Joint Training for Syntactic Analysis*, Université de lille, May 2019, https://tel.archives-ouvertes.fr/tel-02197615

Articles in International Peer-Reviewed Journals

[2] A. Coutant, K. Roper, D. Trejo-Banos, D. Bouthinon, M. Carpenter, J. Grzebyta, G. Santini, H. Soldano, M. Elati, J. Ramon, C. Rouveirol, L. N. Soldatova, R. D. King. *Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast*, in "Proceedings of the National Academy of Sciences of the United States of America ", 2019, vol. 116, n^{o} 36, pp. 18142-18147 [*DOI :* 10.1073/pnas.1900548116], https://hal.sorbonne-universite.fr/hal-02297702

[3] K. Liu, A. Bellet. *Escaping the Curse of Dimensionality in Similarity Learning: Efficient Frank-Wolfe Algorithm and Generalization Bounds*, in "Neurocomputing", 2019, vol. 333, pp. 185-199, https://hal.inria.fr/hal-02166425

[4] A. May, A. Bagheri Garakani, Z. Lu, D. Guo, K. Liu, A. Bellet, L. Fan, M. Collins, D. Hsu, B. Kingsbury, M. Picheny, F. Sha. *Kernel Approximation Methods for Speech Recognition*, in "Journal of Machine Learning Research", 2019, vol. 20, pp. 1 - 36, https://hal.inria.fr/hal-02166422

[15] https://www.arte.tv/fr/videos/RC-017841/dopamine/

[16] https://www.liberation.fr/debats/2019/04/07/grand-debat-et-ia-quelle-transparence-pour-les-donnees_1719944

[17] https://www.editions-ellipses.fr/accueil/4741-apprentissage-machine-cle-de-l-intelligence-artificielle-une-introduction-pour-non-specialistes-9782340028807.html

International Conferences with Proceedings

[5] M. ASADI, M. S. TALEBI, H. BOUREL, O.-A. MAILLARD. *Model-Based Reinforcement Learning Exploiting State-Action Equivalence*, in "ACML 2019, Proceedings of Machine Learning Research", Nagoya, Japan, 2019, vol. 101, pp. 204 - 219, https://hal.archives-ouvertes.fr/hal-02378887

[6] M. BRESSAN, N. CESA-BIANCHI, A. PAUDICE, F. VITALE. *Correlation Clustering with Adaptive Similarity Queries*, in "Conference on Neural Information Processing Systems", Vancouver, Canada, December 2019, https://arxiv.org/abs/1905.11902 , https://hal.inria.fr/hal-02376961

[7] M. DEHOUCK, P. DENIS. *Phylogenetic Multi-Lingual Dependency Parsing*, in "NAACL 2019 - Annual Conference of the North American Chapter of the Association for Computational Linguistics", Minneapolis, United States, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 2019, https://hal.archives-ouvertes.fr/hal-02143747

[8] B. LI, M. DEHOUCK, P. DENIS. *Modal sense classification with task-specific context embeddings*, in "ESANN 2019 - 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning", Bruges, Belgium, April 2019, https://hal.archives-ouvertes.fr/hal-02143762

[9] O. PANDIT, P. DENIS, L. RALAIVOLA. *Learning Rich Event Representations and Interactions for Temporal Relation Classification*, in "ESANN 2019 - 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning", Bruges, Belgium, April 2019, https://hal.archives-ouvertes.fr/hal-02265061

[10] S. PASTERIS, F. VITALE, K. CHAN, S. WANG, M. HERBSTER. *MaxHedge: Maximising a Maximum Online*, in "International Conference on Artificial Intelligence and Statistics", Naha, Okinawa, Japan, April 2019, https://arxiv.org/abs/1810.11843 [*DOI :* 10.11843], https://hal.inria.fr/hal-02376987

[11] B. M. L. SRIVASTAVA, A. BELLET, M. TOMMASI, E. VINCENT. *Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?*, in "INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association", Graz, Austria, September 2019, https://hal.inria.fr/hal-02166434

[12] F. VITALE, A. RAJAGOPALAN, C. GENTILE. *Flattening a Hierarchical Clustering through Active Learning*, in "Conference on Neural Information Processing Systems", Vancouver, Canada, December 2019, https://hal.inria.fr/hal-02376981

[13] R. VOGEL, A. BELLET, S. CLÉMENÇON, O. JELASSI, G. PAPA. *Trade-offs in Large-Scale Distributed Tuplewise Estimation and Learning*, in "ECML PKDD 2019 - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases", Würzburg, Germany, September 2019, https://hal.inria.fr/hal-02166428

Research Reports

[14] A. BELLET, R. GUERRAOUI, H. HENDRIKX. *Who started this rumor? Quantifying the natural differential privacy guarantees of gossip protocols*, Inria, 2019, https://hal.inria.fr/hal-02166432

[15] V. Zantedeschi, A. Bellet, M. Tommasi. *Fully Decentralized Joint Learning of Personalized Models and Collaboration Graphs*, Inria, 2019, https://hal.inria.fr/hal-02166433

Other Publications

[16] J. Bell, A. Bellet, A. Gascón, T. Kulkarni. *Private Protocols for U-Statistics in the Local Model and Beyond*, October 2019, https://arxiv.org/abs/1910.03861 - working paper or preprint [*DOI :* 10.03861], https://hal.inria.fr/hal-02310236

[17] F. Capelli, N. Crosetti, J. Niehren, J. Ramon. *Dependency Weighted Aggregation on Factorized Databases*, January 2019, https://arxiv.org/abs/1901.03633 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01981553

[18] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Ozgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, S. Zhao. *Advances and Open Problems in Federated Learning*, December 2019, https://arxiv.org/abs/1912.04977 - working paper or preprint, https://hal.inria.fr/hal-02406503

[19] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, E. Vincent. *Evaluating Voice Conversion-based Privacy Protection against Informed Attackers*, November 2019, working paper or preprint, https://hal.inria.fr/hal-02355115

[20] W. de Vazelhes, C. Carey, Y. Tang, N. Vauquier, A. Bellet. *metric-learn: Metric Learning Algorithms in Python*, November 2019, https://arxiv.org/abs/1908.04710 - GitHub repository: https://github.com/scikit-learn-contrib/metric-learn, https://hal.inria.fr/hal-02376986

References in notes

[21] A. Alexandrescu, K. Kirchhoff. *Graph-based learning for phonetic classification*, in "IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2007, Kyoto, Japan, December 9-13, 2007", 2007, pp. 359-364

[22] M.-F. Balcan, A. Blum, P. P. Choi, J. Lafferty, B. Pantano, M. R. Rwebangira, X. Zhu. *Person Identification in Webcam Images: An Application of Semi-Supervised Learning*, in "ICML2005 Workshop on Learning with Partially Classified Training Data", 2005

[23] M. Belkin, P. Niyogi. *Towards a Theoretical Foundation for Laplacian-Based Manifold Methods*, in "Journal of Computer and System Sciences", 2008, vol. 74, n^{o} 8, pp. 1289-1308

[24] A. Bellet, A. Habrard, M. Sebban. *A Survey on Metric Learning for Feature Vectors and Structured Data*, in "CoRR", 2013, vol. abs/1306.6709

[25] A. Bellet, A. Habrard, M. Sebban. *Metric Learning*, Morgan & Claypool Publishers, 2015

[26] P. J. BICKEL, A. CHEN. *A nonparametric view of network models and Newman–Girvan and other modularities*, in "Proceedings of the National Academy of Sciences", 2009, vol. 106, pp. 21068–21073

[27] P. BLAU. *Inequality and Heterogeneity: A Primitive Theory of Social Structure*, MACMILLAN Company, 1977, http://books.google.fr/books?id=jvq2AAAAIAAJ

[28] C. BRAUD, P. DENIS. *Combining Natural and Artificial Examples to Improve Implicit Discourse Relation Identification*, in "coling", Dublin, Ireland, August 2014, https://hal.inria.fr/hal-01017151

[29] H. CHANG, D.-Y. YEUNG. *Graph Laplacian Kernels for Object Classification from a Single Example*, in "Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2", Washington, DC, USA, CVPR '06, IEEE Computer Society, 2006, pp. 2011–2016, http://dx.doi.org/10.1109/CVPR.2006.128

[30] D. CHATEL, P. DENIS, M. TOMMASI. *Fast Gaussian Pairwise Constrained Spectral Clustering*, in "ECML/PKDD - 7th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases", Nancy, France, September 2014, pp. 242 - 257 [*DOI :* 10.1007/978-3-662-44848-9_16], https://hal.inria.fr/hal-01017269

[31] D. DAS, S. PETROV. *Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections*, in "ACL", 2011, pp. 600-609

[32] P. DENIS, P. MULLER. *Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition*, in "IJCAI-11 - International Joint Conference on Artificial Intelligence", Barcelone, Espagne, 2011, http://hal.inria.fr/inria-00614765

[33] E. R. FERNANDES, U. BREFELD. *Learning from Partially Annotated Sequences*, in "ECML/PKDD", 2011, pp. 407-422

[34] A. B. GOLDBERG, X. ZHU. *Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization*, in "Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing", Stroudsburg, PA, USA, TextGraphs-1, Association for Computational Linguistics, 2006, pp. 45–52, http://dl.acm.org/citation.cfm?id=1654758.1654769

[35] A. GOLDENBERG, A. X. ZHENG, S. E. FIENBERG. *A Survey of Statistical Network Models*, Foundations and trends in machine learning, Now Publishers, 2010, http://books.google.fr/books?id=gPGgcOf95moC

[36] M. GOMEZ-RODRIGUEZ, J. LESKOVEC, A. KRAUSE. *Inferring networks of diffusion and influence*, in "Proc. of KDD", 2010, pp. 1019-1028

[37] M. MCPHERSON, L. S. LOVIN, J. M. COOK. *Birds of a Feather: Homophily in Social Networks*, in "Annual Review of Sociology", 2001, vol. 27, n^{o} 1, pp. 415–444, http://dx.doi.org/10.1146/annurev.soc.27.1.415

[38] A. NENKOVA, K. MCKEOWN. *A Survey of Text Summarization Techniques*, in "Mining Text Data", Springer, 2012, pp. 43-76

[39] T. RICATTE, R. GILLERON, M. TOMMASI. *Hypernode Graphs for Spectral Learning on Binary Relations over Sets*, in "ECML/PKDD - 7th European Conference on Machine Learning and Principles and Practice

of Knowledge Discovery in Databases", Nancy, France, Machine Learning and Knowledge Discovery in Databases, September 2014, https://hal.inria.fr/hal-01017025

[40] H. SHIN, K. TSUDA, B. SCHÖLKOPF. *Protein functional class prediction with a combined graph*, in "Expert Syst. Appl.", March 2009, vol. 36, n^{o} 2, pp. 3284–3292, http://dx.doi.org/10.1016/j.eswa.2008.01.006

[41] S. SINGH, A. SUBRAMANYA, F. C. N. PEREIRA, A. MCCALLUM. *Large-Scale Cross-Document Coreference Using Distributed Inference and Hierarchical Models*, in "ACL", 2011, pp. 793-803

[42] M. SPERIOSU, N. SUDAN, S. UPADHYAY, J. BALDRIDGE. *Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph*, in "Proceedings of the First Workshop on Unsupervised Methods in NLP", Edinburgh, Scotland, 2011

[43] A. SUBRAMANYA, S. PETROV, F. C. N. PEREIRA. *Efficient Graph-Based Semi-Supervised Learning of Structured Tagging Models*, in "EMNLP", 2010, pp. 167-176

[44] F. VITALE, N. CESA-BIANCHI, C. GENTILE, G. ZAPPELLA. *See the Tree Through the Lines: The Shazoo Algorithm*, in "Proc of NIPS", 2011, pp. 1584-1592

[45] L. WANG, S. N. KIM, T. BALDWIN. *The Utility of Discourse Structure in Identifying Resolved Threads in Technical User Forums*, in "COLING", 2012, pp. 2739-2756

[46] K. K. YUZONG LIU. *Graph-Based Semi-Supervised Learning for Phone and Segment Classification*, in "Proceedings of Interspeech", Lyon, France, 2013

[47] X. ZHU, Z. GHAHRAMANI, J. LAFFERTY. *Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions*, in "Proc. of ICML", 2003, pp. 912-919