IN PARTNERSHIP WITH:
CNRS

Université Charles de Gaulle (Lille 3)

Activity Report 2018

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]
Claudio Gentile [Inria, Senior Researcher, from Oct 2018]

Jan Ramon [Inria, Senior Researcher]

Faculty Members

Bert Cappelle [Univ of Lille, Associate Professor, until Aug 2018]
Mikaela Keller [Univ of Lille, Associate Professor]
Marc Tommasi [Univ of Lille, Team leader, Professor, HDR]
Fabien Torre [Univ of Lille, Associate Professor, until Sep 2018]
Fabio Vitale [Univ of Lille, Associate Professor]
Joël Legrand [Univ of Lille, Associate Professor]

Post-Doctoral Fellows

Melissa Ailem [Inria, until Sep 2018]
Thanh Le Van [Inria, until Jun 2018]
Bo Li [Univ of Lille]

PhD Students

Mahsa Asadi [Inria, from Oct 2018]
Mathieu Dehouck [Univ des sciences et technologies de Lille]
Onkar Pandit [Inria]
Arijus Pleska [Inria, from Oct 2018]
Brij Mohan Lal Srivastava [Inria, from Oct 2018]
Mariana Vargas Vieyra [Inria]

Technical staff

William de Vazelhes [Inria]
Arijus Pleska [Inria, until Sep 2018]
César Sabater [Inria]
Carlos Zubiaga Pena [Inria]

Interns

Igor Axinti [Univ des sciences et technologies de Lille, from Mar 2018 until Aug 2018]
Antoine Capriski [Inria, from May 2018 until Aug 2018]
Arthur d'Azemar [Univ des sciences et technologies de Lille, from Mar 2018 until Aug 2018]
Alexandre Huat [Inria, from Mar 2018 until Aug 2018]

Administrative Assistant

Julie Jonas [Inria]

Visiting Scientist

Tejas Kulkarni [Warwick University, from May 2018 until Aug 2018]

External Collaborator

Remi Gilleron [Univ of 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 on-line 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 [37], [40].

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 [20], [42].

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 Page Rank 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 [41].

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 [20], [45]. 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 [42], [30]. 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 [32].

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 [31], [27], [29]). As a midterm goal, we also plan to work on tasks more directly relating to micro-blogging, such sentiment analysis and the automatic thread structuring of technical forums; the latter task is in fact an instance of rhetorical structure prediction [44]. We have already initiated some work on the coreference resolution with graph-based learning, by casting the problem as an instance of spectral clustering [29].

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 [39], face recognition [28], and text categorization [33].

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 ([46], [21], [22]). 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 ([23], [24]) 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 [35]. 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 [43].

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 [34], 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 [25]. 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 ([26], [36]) 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 [38]. 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. Domain 1

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

- Strengthening of the privacy aware machine learning activity with a new associate team with the Alan Turing Institute and the organization of a workshop at NeurIPS (formerly NIPS).
- New collaboration with Multispeech (Inria Nancy) on decentralized and private machine learning for speech processing leading to an ANR and an H2020 project.

5.1.1. Awards

AURÉLIEN BELLET received a best reviewer award (top 200 out of 3000) at the conference NeurIPS 2018. PASCAL DENIS received a Distinguished Senior Program Committee award at IJCAI-ECAI 2018.

6. New Software and Platforms

6.1. CoRTeX

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.

- 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: William De Vazelhes
- URL: https://github.com/metric-learn/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. On the Bernstein-Hoeffding Method

We consider extensions of Hoeffding's "exponential method" approach for obtaining upper estimates on the probability that a sum of independent and bounded random variables is significantly larger than its mean. We show that the exponential function in Hoeffding's approach can be replaced with any function which is non-negative, increasing and convex. As a result we generalize and improve upon Hoeffding's inequality. Our approach allows to obtain "missing factors" in Hoeffding's inequality. The later result is a rather weaker version of a theorem that is due to Michel Talagrand. Moreover, we characterize the class of functions with respect to which our method yields optimal concentration bounds. Finally, using ideas from the theory of Bernstein polynomials, we show that similar ideas apply under information on higher moments of the random variables ([4]).

7.2. IncGraph: Incremental graphlet counting for topology optimisation

Graphlets are small network patterns that can be counted in order to characterise the structure of a network (topology). As part of a topology optimisation process, one could use graphlet counts to iteratively modify a network and keep track of the graphlet counts, in order to achieve certain topological properties. Up until now, however, graphlets were not suited as a metric for performing topology optimisation; when millions of minor changes are made to the network structure it becomes computationally intractable to recalculate all the graphlet counts for each of the edge modifications. We propose IncGraph, a method for calculating the differences in graphlet counts with respect to the network in its previous state, which is much more efficient than calculating the graphlet occurrences from scratch at every edge modification made. In comparison to static counting approaches, our findings show IncGraph reduces the execution time by several orders of magnitude. The usefulness of this approach was demonstrated by developing a graphlet-based metric to optimise gene regulatory networks. IncGraph is able to quickly quantify the topological impact of small changes to a network, which opens novel research opportunities to study changes in topologies in evolving or online networks, or develop graphlet-based criteria for topology optimisation. IncGraph is freely available as an open-source R package on CRAN (incgraph). The development version is also available on GitHub (rcannood/incgraph) ([2]).

7.3. Graph sampling with applications to estimating the number of pattern embeddings and the parameters of a statistical relational model

Counting the number of times a pattern occurs in a database is a fundamental data mining problem. It is a subroutine in a diverse set of tasks ranging from pattern mining to supervised learning and probabilistic model learning. While a pattern and a database can take many forms, this paper focuses on the case where both the pattern and the database are graphs (networks). Unfortunately, in general, the problem of counting graph occurrences is #P-complete. In contrast to earlier work, which focused on exact counting for simple (i.e., very short) patterns, we present a sampling approach for estimating the statistics of larger graph pattern occurrences. We perform an empirical evaluation on synthetic and real-world data that validates the proposed algorithm, illustrates its practical behavior and provides insight into the trade-off between its accuracy of estimation and computational efficiency ([5]).

7.4. A machine learning based framework to identify and classify long terminal repeat retrotransposons

Transposable elements (TEs) are repetitive nucleotide sequences that make up a large portion of eukaryotic genomes. They can move and duplicate within a genome, increasing genome size and contributing to genetic diversity within and across species. Accurate identification and classification of TEs present in a genome is an important step towards understanding their effects on genes and their role in genome evolution. We introduce TE-LEARNER, a framework based on machine learning that automatically identifies TEs in a given

genome and assigns a classification to them. We present an implementation of our framework towards LTR retrotransposons, a particular type of TEs characterized by having long terminal repeats (LTRs) at their boundaries. We evaluate the predictive performance of our framework on the well-annotated genomes of Drosophila melanogaster and Arabidopsis thaliana and we compare our results for three LTR retrotransposon superfamilies with the results of three widely used methods for TE identification or classification: REPEATMASKER, CENSOR and LTRDIGEST. In contrast to these methods, TE-LEARNER is the first to incorporate machine learning techniques, outperforming these methods in terms of predictive performance, while able to learn models and make predictions efficiently. Moreover, we show that our method was able to identify TEs that none of the above method could find, and we investigated TE-LEARNER's predictions which did not correspond to an official annotation. It turns out that many of these predictions are in fact strongly homologous to a known TE ([6]).

7.5. A Distributed Frank-Wolfe Framework for Learning Low-Rank Matrices with the Trace Norm

We consider the problem of learning a high-dimensional but low-rank matrix from a large-scale dataset distributed over several machines, where low-rankness is enforced by a convex trace norm constraint. We propose DFW-Trace, a distributed Frank-Wolfe algorithm which leverages the low-rank structure of its updates to achieve efficiency in time, memory and communication usage. The step at the heart of DFW-Trace is solved approximately using a distributed version of the power method. We provide a theoretical analysis of the convergence of DFW-Trace, showing that we can ensure sublinear convergence in expectation to an optimal solution with few power iterations per epoch. We implement DFW-Trace in the Apache Spark distributed programming framework and validate the usefulness of our approach on synthetic and real data, including the ImageNet dataset with high-dimensional features extracted from a deep neural network ([7]).

7.6. Personalized and Private Peer-to-Peer Machine Learning

The rise of connected personal devices together with privacy concerns call for machine learning algorithms capable of leveraging the data of a large number of agents to learn personalized models under strong privacy requirements. In this paper, we introduce an efficient algorithm to address the above problem in a fully decentralized (peer-to-peer) and asynchronous fashion, with provable convergence rate. We show how to make the algorithm differentially private to protect against the disclosure of information about the personal datasets, and formally analyze the trade-off between utility and privacy. Our experiments show that our approach dramatically outperforms previous work in the non-private case, and that under privacy constraints, we can significantly improve over models learned in isolation ([9]).

7.7. Hiding in the Crowd: A Massively Distributed Algorithm for Private Averaging with Malicious Adversaries

The amount of personal data collected in our everyday interactions with connected devices offers great opportunities for innovative services fueled by machine learning, as well as raises serious concerns for the privacy of individuals. In this paper, we propose a massively distributed protocol for a large set of users to privately compute averages over their joint data, which can then be used to learn predictive models. Our protocol can find a solution of arbitrary accuracy, does not rely on a third party and preserves the privacy of users throughout the execution in both the honest-but-curious and malicious adversary models. Specifically, we prove that the information observed by the adversary (the set of malicious users) does not significantly reduce the uncertainty in its prediction of private values compared to its prior belief. The level of privacy protection depends on a quantity related to the Laplacian matrix of the network graph and generally improves with the size of the graph. Furthermore, we design a verification procedure which offers protection against malicious users joining the service with the goal of manipulating the outcome of the algorithm ([15]).

7.8. A Probabilistic Model for Joint Learning of Word Embeddings from Texts and Images

Several recent studies have shown the benefits of combining language and perception to infer word embeddings. These multimodal approaches either simply combine pre-trained textual and visual representations (e.g. features extracted from convolutional neural networks), or use the latter to bias the learning of textual word embeddings. In this work, we propose a novel probabilistic model to formalize how linguistic and perceptual inputs can work in concert to explain the observed word-context pairs in a text corpus. Our approach learns textual and visual representations jointly: latent visual factors couple together a skip-gram model for co-occurrence in linguistic data and a generative latent variable model for visual data. Extensive experimental studies validate the proposed model. Concretely, on the tasks of assessing pairwise word similarity and image/caption retrieval, our approach attains equally competitive or stronger results when compared to other state-of-the-art multimodal models ([8]).

7.9. A Framework for Understanding the Role of Morphology in Universal Dependency Parsing

We present a simple framework for characterizing morphological complexity and how it encodes syntactic information. In particular, we propose a new measure of morpho-syntactic complexity in terms of governor-dependent preferential attachment that explains parsing performance. Through experiments on dependency parsing with data from Universal Dependencies (UD), we show that representations derived from morphological attributes deliver important parsing performance improvements over standard word form embeddings when trained on the same datasets. We also show that the new morpho-syntactic complexity measure is predictive of the gains provided by using morphological attributes over plain forms on parsing scores, making it a tool to distinguish languages using morphology as a syntactic marker from others ([11]).

7.10. Online Reciprocal Recommendation with Theoretical Performance Guarantees

A reciprocal recommendation problem is one where the goal of learning is not just to predict a user's preference towards a passive item (e.g., a book), but to recommend the targeted user on one side another user from the other side such that a mutual interest between the two exists. The problem thus is sharply different from the more traditional items-to-users recommendation, since a good match requires meeting the preferences at both sides. We initiate a rigorous theoretical investigation of the reciprocal recommendation task in a specific framework of sequential learning. We point out general limitations, formulate reasonable assumptions enabling effective learning and, under these assumptions, we design and analyze a computationally efficient algorithm that uncovers mutual likes at a pace comparable to that achieved by a clairvoyant algorithm knowing all user preferences in advance. Finally, we validate our algorithm against synthetic and real-world datasets, showing improved empirical performance over simple baselines ([13]).

7.11. On Similarity Prediction and Pairwise Clustering

We consider the problem of clustering a finite set of items from pairwise similarity information. Unlike what is done in the literature on this subject, we do so in a passive learning setting, and with no specific constraints on the cluster shapes other than their size. We investigate the problem in different settings: i. an online setting, where we provide a tight characterization of the prediction complexity in the mistake bound model, and ii. a standard stochastic batch setting, where we give tight upper and lower bounds on the achievable generalization error. Prediction performance is measured both in terms of the ability to recover the similarity function encoding the hidden clustering and in terms of how well we classify each item within the set. The proposed algorithms are time efficient ([12]).

7.12. A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization

The performance of many machine learning techniques depends on the choice of an appropriate similarity or distance measure on the input space. Similarity learning (or metric learning) aims at building such a measure from training data so that observations with the same (resp. different) label are as close (resp. far) as possible. In this paper, similarity learning is investigated from the perspective of pairwise bipartite ranking, where the goal is to rank the elements of a database by decreasing order of the probability that they share the same label with some query data point, based on the similarity scores. A natural performance criterion in this setting is pointwise ROC optimization: maximize the true positive rate under a fixed false positive rate. We study this novel perspective on similarity learning through a rigorous probabilistic framework. The empirical version of the problem gives rise to a constrained optimization formulation involving U-statistics, for which we derive universal learning rates as well as faster rates under a noise assumption on the data distribution. We also address the large-scale setting by analyzing the effect of sampling-based approximations. Our theoretical results are supported by illustrative numerical experiments ([14]).

7.13. 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 this paper, 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 ([16]).

7.14. Nonstochastic Bandits with Composite Anonymous Feedback

We investigate a nonstochastic bandit setting in which the loss of an action is not immediately charged to the player, but rather spread over at most d consecutive steps in an adversarial way. This implies that the instantaneous loss observed by the player at the end of each round is a sum of as many as d loss components of previously played actions. Hence, unlike the standard bandit setting with delayed feedback, here the player cannot observe the individual delayed losses, but only their sum. Our main contribution is a general reduction transforming a standard bandit algorithm into one that can operate in this harder setting. We also show how the regret of the transformed algorithm can be bounded in terms of the regret of the original algorithm. Our reduction cannot be improved in general: we prove a lower bound on the regret of any bandit algorithm in this setting that matches (up to log factors) the upper bound obtained via our reduction. Finally, we show how our reduction can be extended to more complex bandit settings, such as combinatorial linear bandits and online bandit convex optimization ([10]).

8. Bilateral Contracts and Grants with Industry

8.1. Coreference resolution

Along a collaboration with Orange, we developed a Natural Language Processing library for co-reference resolution. The library is based on a previous work (CorTeX) and was extended in several ways. It handles the French language, it includes new features based on vectorial representations of words (word embeddings) and it is more scalable. PASCAL DENIS is the local PI at Inria of this project.

8.2. Privacy preserving data mining for Mobility Data

JAN RAMON is the local PI at Inria for the ADEME-MUST project (Méthodologie d'exploitation des données d'usage des véhicules et d'identification de nouveaux services pour les usagers et les territoires). We study machine learning and data mining methods for knowledge discovery from mobility data, which are time-stamped signals collected from cars, for example, GPS locations, accelerations and fuel consumption. We aim to discover knowledge that helps us to address important questions in the transportation system such as road safety, traffic congestion, parking, ride-sharing, pollution and energy consumption. As the mobility data contains a lot of personal information, for instance, driving styles and locations of the users, we hence also study methods that allow the users to keep their personal data and only exchange part of them to collaboratively derive the knowledge.

The project has four partners, including, Xee company, CEREMA, i-Trans and Inria. The Xee company is responsible for recruiting drivers and collecting the data. CEREMA and i-Trans function as domain experts who help us to form the questions and verify the analytical results. MAGNET is responsible for developing and applying data mining methods for analyzing the data. The developed methods and the discovered knowledge from the project will be transferred to Metropole Lille and ADEME.

8.3. Predictive justice

Claim assistance is a French company that develops assistance for conflict resolution. The main service is RefundMyTicket [1]. In the general project of partial automation of analysis of complains, we have provided consulting and supervision. The general approach was to be able to analyze, parse and reason on legal texts. We have developed strategies based on natural language processing in the specific domain of legal texts. Techniques include learning representation and structured prediction among others.

9. Partnerships and Cooperations

9.1. Regional Initiatives

We conducted research in collaboration with J. Senechal from the department of law in Lille University. We are interested in studying the impact of technological choices regarding computation models in the perspective of the GDPR.

We strengthened our partnership with the linguistic laboratory STL in Lille university. We have welcomed Bert Cappelle for a stay (delegation) in the group. The topic of this collaboration was to study modal verbs and the translation of the notion of compositionality when applied to vectorial representation of words.

We initiated a collaboration with cognitive scientists (Angèle Brunellière and Jérémie Jozefowiez) from the psychology department, which resulted in a submission to a multidisciplinary Huma-Num project, to be funded by the Réseau National des Maisons des Sciences de l'Homme (RNMSH).

We started working with Christopher Fletcher (CNRS) from the History department.

These collaborations heavily rely on our work on distributional semantics and word embeddings to provide new insights into these different fields, hence also on the Mangoes toolkit developed in the team.

[1] https://www.refundmyticket.net

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.

9.2. National Initiatives

9.2.1. 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/.

9.2.2. 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/.

9.2.3. 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.

9.2.4. ANR-NFS REM (2016-2020)

Participants: PASCAL DENIS [correspondent], BO LI

With colleagues from the linguistics departments at Lille 3 and 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 (Lille 3) 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.

9.2.5. EFL (2010-2020)

PASCAL DENIS is an associate member of the Laboratoire d'Excellence *Empirical Foundations of Linguistics* (EFL), http://www.labex-efl.org/.

9.3. European Initiatives

9.3.1. FP7 & H2020 Projects

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

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

Participants: AURÉLIEN BELLET, MARC TOMMASI, BRIJ SRIVASTAVA

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.

9.3.2. Collaborations in European Programs, Except FP7 & H2020

9.3.2.1. TextLink (2014-2018)

Program: COST Action

Project acronym: TextLink

Project title: Structuring Discourse in Multilingual Europe

Duration: Apr. 2014 - Apr. 2018

Coordinator: Prof. Liesbeth Degand, Université Catholique de Louvain, Belgium. PASCAL DENIS is member of the Tools group.

Other partners: 26 EU countries and 3 international partner countries (Argentina, Brazil, Canada)

The Action will facilitate European multilingualism by (1) identifying and creating a portal into such resources within Europe - including annotation tools, search tools, and discourse-annotated corpora; (2) delineating the dimensions and properties of discourse annotation across corpora; (3) organizing these properties into a sharable taxonomy; (4) encouraging the use of this taxonomy in subsequent discourse annotation and in cross-lingual search and studies of devices that relate and structure discourse; and (5) promoting use of the portal, its resources and sharable taxonomy. TextLink will enhance the experience and performance of human translators, lexicographers, language technology and language learners alike.

9.4. International Initiatives

9.4.1. Inria International Labs

Inria@SiliconValley
Associate Team involved in the International Lab:

9.4.1.1. LEGO

Title: LEarning GOod representations for natural language processing
International Partner (Institution - Laboratory - Researcher):
USC (United States), Prof. Fei Sha.
Start year: 2016
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 structured data and how to learn them automatically, and how to apply such representations to complex and structured prediction tasks in NLP? In recent years, continuous vectorial embeddings learned from massive unannotated corpora have been increasingly popular, but they remain far too limited to capture the complexity of text data as they are task-agnostic and fall short of modeling complex structures in languages. LEGO strongly relies on the complementary expertise of the two partners in areas such as representation/similarity learning, structured prediction, graph-based learning, and statistical NLP to offer a novel alternative to existing techniques. Specifically, we will investigate the following three research directions: (a) optimize the embeddings based on annotations so as to minimize structured prediction errors, (b) generate embeddings from rich language contexts represented as graphs, and (c) automatically adapt the context graph to the task/dataset of interest by learning a similarity between nodes to appropriately weigh the edges of the graph. By exploring these complementary research strands, we intend to push the state-of-the-art in several core NLP problems, such as dependency parsing, coreference resolution and discourse parsing.

9.4.2. 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.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

- Tejas Kulkarni (University of Warwick) visited the team from May to August 2018 to work with AURÉLIEN BELLET, MARC TOMMASI and JAN RAMON on privacy-preserving computation of U-statistics.
- Larisa Soldatova (Brunel University) visited the team in June 2018 to work with JAN RAMON on probabilistic reasoning for biomedical applications.
- Raouf Kerkouche (Inria Privatics) visited the team for 2 weeks in July 2018 to work with AURÉLIEN BELLET and MARC TOMMASI on federated and decentralized learning from medical data.
- Guillaume Rabusseau (Université de Montréal) visited the team for 1 week in July 2018 to work with AURÉLIEN BELLET and MARC TOMMASI on multi-task distributed spectral learning.
- Daphner Ezer, Adrià Gascón, Matt Kusner, Brooks Paige (all from Alan Turing Institute) and Hamed Haddadi (Imperial College London) visited the team for 2 days in October 2018 for the kick-off of the PAD-ML associate team.

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

- D. Hovy (Bocconi Univ.): Retrofit Everything: Injecting External Knowledge into Neural Networks to Gain Insights from Big Data.
- A. Trask (OpenMined): OpenMined - Building Tools for Safe AI.
- C. Biemann (Univ. Hamburg): Adaptive Interpretable Language Technology.
- W. Daelemans (Univ. Antwerp): Profiling authors from social media texts.

9.5.1.1. Internships

- Igor Axinti explored several ways to compare word embeddings and studied the minimal corpus size for the comparison to be meaningful. He applied some of his findings to comparing two corpus in middle french from the 15th century, one originating from London and the other from Flanders. He produced a querying interface to allow Christopher Fletcher (IRHiS), who provided the data, explore and compare the embeddings spaces.
- Nicolas Crosetti (joint internship with Joachim Niehren and Florent Cappelli, Links) worked on dependency-weighted aggregation, i.e., aggregation where the elements to aggregate are weighted according to the extent where they correspond to independent observations.
- Arthur d'Azemar worked on decentralized recommender systems in collaboration with the WIDE team in Inria Rennes (François Taïani). Arthur has applied metric learning techniques in order to learn a K-nn graph for personalized and adaptive user-based recommendations.
- Antoine Capriski worked on the analysis of word semantic change in political texts in collaboration with Caroline Le Pennec (UC Berkeley). He used the techniques of word embeddings to analyze of corpus of political manifestos from the French general elections for the period 1958-1993.
- Most of the works on machine learning and privacy make the assumption that learners are honest but curious. Alexandre Huat worked on making protocols for private machine learning more robust again malicious attacks.

9.5.2. Visits to International Teams

9.5.2.1. Research Stays Abroad

- FABIO VITALE is 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 follows three directions:
  - designing new online reciprocal recommenders analyzing their performance both in theory and in practice,
  - clustering a finite set of items from pairwise similarity information in different learning settings,
  - introducing a new online learning framework encompassing several problems where the environment changes over time, and an efficient and very scalable unifying approach to solve the related general learning problem.

  Current (and unfinished) ongoing research also includes the following topics: low-stretch spanning trees, active learning in correlation clustering problems, hierarchical clustering.
- AURÉLIEN BELLET visited the Alan Turing Institute (London) and Amazon Research Cambridge for 1 week in February 2018. He worked with Adrià Gascón and Borja Balle on privacy-preserving machine learning.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. Member of the Organizing Committees

- AURÉLIEN BELLET was a member of the organization committee of the PPML workshop at NeurIPS'18. [2] The workshop was on Privacy Preserving Machine Learning and had among its invited speakers Shafi Goldwasser (Gödel and Turing Prize), Adam Smith (Gödel Prize).
- AURÉLIEN BELLET co-organized the kick-off workshop of the associated team PAD-ML with the Alan Turing Institute. [3] The workshop was held at Inria Lille and featured speakers from MAGNET and the Alan Turing Institute.

[2] https://neurips.cc/Conferences/2018/Schedule?showEvent=10934
[3] https://team.inria.fr/magnet/workshop-on-privacy-aware-distributed-machine-learning/

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

- AURÉLIEN BELLET served as PC member for AISTATS'19, ICML'18, NIPS'18, IJCAI'18 Sister Conference, PiMLAI workshop at ICML'18, and CAP'18.
- PASCAL DENIS served as PC member for ACL'18, CONLL'18, EMNLP'18, NAACL'18, NIPS'18, IJCAI-ECAI'18 (Senior PC), CRAC Workshop at NAACL'18.
- MARC TOMMASI served as PC member for AAAI'18, ICML'18, CAP'18, IJCAI'18 (Senior PC chair), AISTATS'18, NIPS'18.
- JAN RAMON served as PC member for AAAI'19, AISTATS'19, IEEE-BigData'18, CIKM'18, DS'18, ECML/PKDD'18, EKAW'18, IEEE-ICDM'18, ICML'18, ILP'18, LOD'18, MLG'18, NIPS'18, SDM'18, TDLSG'18.
- MIKAELA KELLER served as PC member for ICML'18, CAP'18.
- RÉMI GILLERON served as PC member for NIPS'18, CAP'18, AISTATS'19 and ICLR'19.

10.1.3. Journal

10.1.3.1. Reviewer - Reviewing Activities

- AURÉLIEN BELLET was reviewer for Machine Learning Journal and IEEE/ACM Transactions on Networking.
- PASCAL DENIS was reviewer for Computational Linguistics, IJCAI-ECAI Surveys, and Language Resources and Evaluation.
- JAN RAMON was member of the editorial boards of Machine Learning Journal (MLJ) and Data Mining and Knowledge Discovery (DMKD). JAN RAMON was reviewer for among others JMLR, TPAMI, JIIS.

10.1.4. Invited Talks

- AURÉLIEN BELLET gave invited talks at the EPFL-Inria 2018 workshop [4] and the Journées de Statistique 2018 (session SSFAM). [5]
- AURÉLIEN BELLET was invited to talk at the seminars of Inria WIDE, Télécom ParisTech, Statistics Seminar of Paris 6/7, CMLA (ENS Paris Saclay) and Naver Labs Europe.
- PASCAL DENIS gave an invited talk at the Séminaire Langage, SCALab, Université de Lille, 26/01/18.

10.1.5. 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. [6]
- AURÉLIEN BELLET acted as external reviewer for the French National Research Agency (ANR), track "Projets de Recherche Collaborative – International".
- JAN RAMON was an external reviewer for the Swiss National Science Foundation (SNF).
- JAN RAMON was an external reviewer for the Vienna Science and Technology Fund (WWTF).
- JAN RAMON acted as an expert for the H2020 CoE and IMI programs.

[4] https://project.inria.fr/epfl-Inria/workshops/workshop-2018/
[5] http://jds2018.sfds.asso.fr/
[6] https://www.societe-informatique-de-france.fr/recherche/prix-de-these-gilles-kahn/

10.1.6. Research Administration

- MIKAELA KELLER is member of the Conseil du laboratoire CRIStAL.
- FABIEN TORRE is member of the bureau du Conseil National des Universités (section 27).
- PASCAL DENIS served as a member of the CNRS Pre-GDR NLP Group.
- PASCAL DENIS was elected to Comité National du CNRS, section 34 (Sciences du Langage).

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Licence SHS: JOËL LEGRAND, Traitement de textes et tableur, 10h, L1, Université Lille.

Licence SHS: MARC TOMMASI, Langages du Web, 24h, L2, Université Lille.

Licence MIASHS: MIKAELA KELLER, Python 1, 40h, L1, Université Lille.

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

Licence MIASHS: MIKAELA KELLER, Codage et représentation de l'information, 42h, L1, Université Lille.

Licence SoQ (SHS): MIKAELA KELLER, Algorithmique de graphes, 24h, L3, Université Lille.

Licence MARC TOMMASI C2i 12h, Université Lille.

Licence MARC TOMMASI Humanités numériques - Découvrir et faire découvrir la programmation, 20h, Université Lille/

Master MIASHS: MIKAELA KELLER, Algorithmes fondamentaux de la fouille de données, 60h, M1, Université Lille.

Master MIASHS: JOËL LEGRAND, Apprentissage et émergence de comportements, 30h, M2, Université Lille.

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, 17.5h, Télécom ParisTech.

Master Informatique: PASCAL DENIS, Fondements de l'Apprentissage Automatique, 46h, M1, Université de Lille.

10.2.2. Supervision

Postdoc: MELISSA AILEM, InriaSiliconValley postdoctoral grant, supervised by AURÉLIEN BELLET, MARC TOMMASI, PASCAL DENIS and FEI SHA (University of Southern California).

Postdoc: BO LI, supervised by PASCAL DENIS on ANR REM, Model Sense Disambiguation, since December 2017.

PhD: GÉRAUD LE FALHER, Characterizing edges in signed and vector-valued graphs. April 16th 2018, MARC TOMMASI and FABIO VITALE and CLAUDIO GENTILE.

Phd: ASHRAF M. KIBRIYA, Mining Frequent Patterns in Large Networks, June 2018, JAN RAMON.

PhD in progress: MATHIEU DEHOUCK, Graph-based Learning for Multi-lingual and Multi-domain Dependency Parsing, since Oct 2015, 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.

10.2.3. Juries

- AURÉLIEN BELLET was member of the PhD jury of Guillaume Papa (Télécom ParisTech), Wenjie Zheng (Sorbonne Université), Michael Blot (Sorbonne Université).
- MARC TOMMASI was member of the Phd jury of Gaëtan Hadjeres (*Rapporteur*), Alexandre Bérard (*Head*), Olivier Ruas (*Rapporteur*), Valentina Zantedeschi.
- PASCAL DENIS was *rapporteur* on the Phd jury of Elena Knyazeva, Université Paris-Saclay.
- MIKAELA KELLER was member of the recruitment committee for Assistant Professors in Computer Science at Université of Lille and at Université de St-Étienne.
- MIKAELA KELLER was member of the Phd jury of Damien Fourure (Université de St-Étienne) and of the HDR jury of Renaud Lopes (CHRU Lille).
- RÉMI GILLERON was head of the PhD jury of Romain Warlop (Université de Lille).
- PASCAL DENIS was a member of hiring committee for Junior Research Scientist at Inria Lille.
- MARC TOMMASI was member of the recruitment committee Assistant Professors in Computer Science at Université of Lille and for professor position at INSA de Lyon.

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

- AURÉLIEN BELLET is the scientific mediation contact for Inria Lille center.
- PASCAL DENIS served as committee member on the Inria Lille Commission Emploi Recherche (CER).
- PASCAL DENIS also served as committee member on Commission de Développement Technologique (CDT).
- PASCAL DENIS is administrator of Inria membership to Linguistic Data Consortium (LDC).

10.3.2. Articles and contents

- AURÉLIEN BELLET and MARC TOMMASI provided expertise for an upcoming TV program on Arte about new technologies.

10.3.3. Interventions

- National events: JAN RAMON and MARC TOMMASI participate to a round-table meeting at the *Fête des libertés numériques* for the RGPD day [7].
- In educational institutions: MARC TOMMASI gave a talk on privacy and machine learning in Journées polytech [8].

[7] https://www.meshs.fr/page/donnees_personnelles_et_droits_et_libertes_numeriques
[8] http://www.polytech-lille.fr/big-data-machine-mearning-p11419.html#.WqlHjExFxPb

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] G. LE FALHER. *Characterizing Edges in Signed and Vector-Valued Graphs*, Université de Lille, April 2018, https://hal.inria.fr/tel-01824215

Articles in International Peer-Reviewed Journals

[2] R. CANNOODT, J. RUYSSINCK, J. RAMON, K. DE PRETER, Y. SAEYS. *IncGraph: Incremental graphlet counting for topology optimisation*, in "PLoS ONE", April 2018, vol. 13, n° 4 [*DOI :* 10.1371/ JOURNAL.PONE.0195997], https://hal.inria.fr/hal-01814675

[3] B. CAPPELLE, P. DENIS, M. KELLER. *Facing the facts of fake: a distributional semantics and corpus annotation approach*, in "Yearbook of the German Cognitive Linguistics Association", November 2018, https://hal.archives-ouvertes.fr/hal-01959609

[4] C. PELEKIS, J. RAMON, Y. WANG. *On the Bernstein-Hoeffding method*, in "Bulletin of the Hellenic Mathematical Society", June 2018, vol. 62, pp. 31-43, https://hal.inria.fr/hal-01814651

[5] I. RAVKIC, M. ZNIDARŠIČ, J. RAMON, J. DAVIS. *Graph sampling with applications to estimating the number of pattern embeddings and the parameters of a statistical relational model*, in "Data Mining and Knowledge Discovery", 2018, 36 p. , https://hal.inria.fr/hal-01725971

[6] L. SCHIETGAT, C. VENS, J. RAMON, R. CERRI, C. N. FISCHER, E. D. COSTA, C. M. A. CARARETO, H. BLOCKEEL. *A machine learning based framework to identify and classify long terminal repeat retrotransposons*, in "PLoS Computational Biology", April 2018, vol. 14, n° 4, pp. 1-21 [*DOI :* 10.1371/JOURNAL.PCBI.1006097], https://hal.inria.fr/hal-01814669

[7] W. ZHENG, A. BELLET, P. GALLINARI. *A Distributed Frank-Wolfe Framework for Learning Low-Rank Matrices with the Trace Norm*, in "Machine Learning", 2018, https://hal.inria.fr/hal-01922994

International Conferences with Proceedings

[8] M. AILEM, B. ZHANG, A. BELLET, P. DENIS, F. SHA. *A Probabilistic Model for Joint Learning of Word Embeddings from Texts and Images*, in "Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)", Brussels, Belgium, 2018, https://hal.inria.fr/hal-01922985

[9] A. BELLET, R. GUERRAOUI, M. TAZIKI, M. TOMMASI. *Personalized and Private Peer-to-Peer Machine Learning*, in "AISTATS 2018 - 21st International Conference on Artificial Intelligence and Statistics", Lanzarote, Spain, April 2018, pp. 1-20, https://arxiv.org/abs/1705.08435 , https://hal.inria.fr/hal-01745796

[10] N. CESA-BIANCHI, C. GENTILE, Y. MANSOUR. *Nonstochastic Bandits with Composite Anonymous Feedback*, in "COLT 2018 - 31st Annual Conference on Learning Theory", Stockholm, Sweden, July 2018, vol. 75, pp. 1 - 23, https://hal.inria.fr/hal-01916981

[11] M. DEHOUCK, P. DENIS. *A Framework for Understanding the Role of Morphology in Universal Dependency Parsing*, in "EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing", Brussels, Belgium, Proceedings of EMNLP 2018, October 2018, https://hal.archives-ouvertes.fr/hal-01943934

[12] S. PASTERIS, F. VITALE, C. GENTILE, M. HERBSTER. *On Similarity Prediction and Pairwise Clustering*, in "ALT 2018 - 29th International Conference on Algorithmic Learning Theory", Lanzarote, Spain, April 2018, vol. 83, pp. 1 - 28, https://hal.inria.fr/hal-01916976

[13] F. VITALE, N. PAROTSIDIS, C. GENTILE. *Online Reciprocal Recommendation with Theoretical Performance Guarantees*, in "NIPS 2018 - 32nd Conference on Neural Information Processing Systems", Montreal, Canada, December 2018, https://hal.inria.fr/hal-01916979

[14] R. VOGEL, A. BELLET, S. CLÉMENÇON. *A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization*, in "International Conference on Machine Learning (ICML 2018)", Stockholm, Sweden, 2018, https://hal.inria.fr/hal-01922988

Research Reports

[15] P. DELLENBACH, A. BELLET, J. RAMON. *Hiding in the Crowd: A Massively Distributed Algorithm for Private Averaging with Malicious Adversaries*, Inria, 2018, https://hal.inria.fr/hal-01923000

[16] K. LIU, A. BELLET. *Escaping the Curse of Dimensionality in Similarity Learning: Efficient Frank-Wolfe Algorithm and Generalization Bounds*, Inria, 2018, https://hal.inria.fr/hal-01923006

Other Publications

[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] O.-A. MAILLARD, M. ASADI. *Upper Confidence Reinforcement Learning exploiting state-action equivalence*, December 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01945034

[19] J. RAMON. *Exploiting traffic data*, December 2018, Presentation given at Inria Meetup (Lille), https://hal.inria.fr/hal-01879941

References in notes

[20] 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

[21] 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

[22] 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

[23] A. BELLET, A. HABRARD, M. SEBBAN. *A Survey on Metric Learning for Feature Vectors and Structured Data*, in "CoRR", 2013, vol. abs/1306.6709

[24] A. BELLET, A. HABRARD, M. SEBBAN. *Metric Learning*, Morgan & Claypool Publishers, 2015

[25] 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

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

[27] 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

[28] 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

[29] 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

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

[31] 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

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

[33] 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

[34] 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

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

[36] 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

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

[38] 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

[39] 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

[40] 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

[41] 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

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

[43] 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

[44] 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

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

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