IN PARTNERSHIP WITH:
CNRS

Université Charles de Gaulle (Lille 3)

Activity Report 2017

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
B1.1.11. - Systems 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
B9.4.1. - Computer science
B9.4.4. - Chemistry
B9.4.5. - Data science
B9.5.8. - Linguistics
B9.5.10. - Digital humanities
B9.8. - Privacy

1. Personnel

Research Scientists

Aurelien Bellet [Inria, Researcher]
Pascal Denis [Inria, Researcher]
Claudio Gentile [Inria, Advanced Research Position, from Nov 2017]

Jan Ramon [Inria, Senior Researcher]

Faculty Members

Marc Tommasi [Team leader, Univ Charles de Gaulle, Professor, HDR]
Remi Gilleron [Univ Charles de Gaulle, Professor, HDR]
Mikaela Keller [Univ Charles de Gaulle, Associate Professor]
Fabien Torre [Univ Charles de Gaulle, Associate Professor]
Fabio Vitale [Univ Charles de Gaulle, Associate Professor]

Post-Doctoral Fellows

Melissa Ailem [Inria, from Oct 2017]
Thanh Le Van [Inria, from Mar 2017]
Bo Li [Univ Charles de Gaulle, from Dec 2017]

PhD Students

Mathieu Dehouck [Univ des sciences et technologies de Lille]
Geraud Le Falher [Inria, until Nov 2017]
Thibault Lietard [Univ Charles de Gaulle]
Onkar Pandit [Inria, from Dec 2017]

Technical staff

William de Vazelhes [Inria, from Sep 2017]
Arijus Pleska [Inria, from Nov 2017]
Carlos Zubiaga Pena [Inria, from Jun 2017]

Interns

Hippolyte Bourel [Inria, from May 2017 until Jul 2017]
Juhi Tandon [Univ des sciences et technologies de Lille, from Jun 2017 until Aug 2017]
Quentin Tremouille [Inria, until Apr 2017]

Administrative Assistant

Julie Jonas [Inria]

Visiting Scientists

Wilhelmiina Hamalainen [Aalto University, from Feb 2017 until Mar 2017]
Peter Kling [Hamburg University, Sep 2017]
Clement Weisbecker [Livermore Software Technology Corporation, from Aug 2017 until Sep 2017]
Valentina Zantedeschi [Univ. St Etienne, Sep 2017]
Isabel Valera [Max Planck Institute, Nov 2017]
Bert Cappelle [Univ Charles de Gaulle, from Sep 2017]

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 [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 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 [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 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. Domain

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

- First public release of the Mangoes software
- Major publications in machine learning and natural language processing (AISTATS'17, EACL'17)
- Increased visibility of the team on decentralized learning and privacy with applications on mobility data through publications, Workshops, invited talks and bilateral contracts
- Participation in the the Scikit-Learn development team with AURÉLIEN BELLET and WILLIAM DE VAZELHES through the ADT SkMetricLearn

5.1.1. Awards

- AURÉLIEN BELLET was awarded the Prime d'encadrement doctoral et de recherche (PEDR), category "junior"
- PASCAL DENIS was awarded the Prime d'encadrement doctoral et de recherche (PEDR), category "confirmé"

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 main annotation formats (ACE, CoNLL and MUC), and performing evaluation 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.

- Participant: Pascal Denis
- Contact: Pascal Denis
- URL: https://gforge.inria.fr/projects/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

7. New Results

7.1. Natural Language Processing

In [13] we present a new, efficient method for learning task-specific word vectors using a variant of the Passive-Aggressive algorithm. Specifically, this algorithm learns a word embedding matrix in tandem with the classifier parameters in an online fashion, solving a bi-convex constrained optimization at each iteration. We provide a theoretical analysis of this new algorithm in terms of regret bounds, and evaluate it on both synthetic data and NLP classification problems, including text classification and sentiment analysis. In the latter case, we compare various pre-trained word vectors to initialize our word embedding matrix, and show that the matrix learned by our algorithm vastly outperforms the initial matrix, with performance results comparable or above the state-of-the-art on these tasks.

In [12] we present a new approach to the problem of cross-lingual dependency parsing, aiming at leveraging training data from different source languages to learn a parser in a target language. Specifically, this approach first constructs word vector representations that exploit structural (i.e., dependency-based) contexts but only considering the morpho-syntactic information associated with each word and its contexts. These delexicalized word em-beddings, which can be trained on any set of languages and capture features shared across languages, are then used in combination with standard language-specific features to train a lexicalized parser in the target language. We evaluate our approach through experiments on a set of eight different languages that are part the Universal Dependencies Project. Our main results show that using such delexicalized embeddings, either trained in a monolingual or multilingual fashion, achieves significant improvements over monolingual baselines.

7.2. Decentralized Learning and Privacy

In [15] we consider a set of learning agents in a collaborative peer-to-peer network, where each agent learns a personalized model according to its own learning objective. The question addressed in this paper is: how can agents improve upon their locally trained model by communicating with other agents that have similar objectives? We introduce and analyze two asynchronous gossip algorithms running in a fully decentralized manner. Our first approach, inspired from label propagation, aims to smooth pre-trained local models over the network while accounting for the confidence that each agent has in its initial model. In our second approach, agents jointly learn and propagate their model by making iterative updates based on both their local dataset and the behavior of their neighbors. To optimize this challenging objective, our decentralized algorithm is based on ADMM.

In subsequent work in collaboration with Rachid Guerraoui's group at EPFL [18], we study how agents can collaborate to improve upon their locally learned model without leaking sensitive information about their data. Our first contribution is to reformulate this problem so that it can be solved by a block coordinate descent algorithm. We obtain an efficient and fully decentralized protocol working in an asynchronous fashion. Our second contribution is to make our algorithm differentially private to protect against the disclosure of any information about personal datasets. We prove convergence rates and exhibit 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 significantly improve over purely local models. A preliminary version of this work was presented at the NIPS 2017 workshop on machine Learning on the Phone and other Consumer Devices [16].

7.3. Statistical Learning on Graphs

The main purpose of [11] is to illustrate that certain Hölder-type inequalities can be employed in order to obtain concentration and correlation bounds for sums of weakly dependent random variables whose dependencies are described in terms of graphs, or hypergraphs. Let Y_v , $v \in V$, be real-valued random variables having a dependency graph $G = (V, E)$. We show that

$$\mathbb{E}\left[\prod_{v\in V} Y_v\right] \leq \prod \left\{\mathbb{E}\left[Y_v^{\frac{\chi_b}{b}}\right]\right\}^{\frac{b}{\chi_b}}$$

where χ_b is the b-fold chromatic number of G. This inequality may be seen as a dependency-graph analogue of a generalized Hölder inequality, due to Helmut Finner. Additionally, we provide applications of the afore-mentioned Hölder-type inequalities to concentration and correlation bounds for sums of weakly dependent random variables whose dependencies can be described in terms of graphs or hypergraphs.

Several collaborations concerned efficient counting of subgraph frequencies in networks. Two journal articles are accepted subject to minor revisions, one in collaboration with the group of Yvan Saeys (University of Ghent, Belgium), and one in collaboration with Irma Ravkic and Martin Znidarsic (former collaborators of JAN RAMON).

7.4. Data Mining with Rank Data

Rank data, in which each row is a complete or partial ranking of available items (columns), is ubiquitous. Among others, it can be used to represent preferences of users, levels of gene expression, and outcomes of sports events. It can have many types of patterns, among which consistent rankings of a subset of the items in multiple rows, and multiple rows that rank the same subset of the items highly. In [10], we show that the problems of finding such patterns can be formulated within a single generic framework that is based on the concept of semiring matrix factorization. In this framework, we employ the max-product semiring rather than the plus-product semiring common in traditional linear algebra. We apply this semiring matrix factorization framework on two tasks: sparse rank matrix factorization and rank matrix tiling. Experiments on both synthetic and real world datasets show that the framework is capable of discovering different types of structure as well as obtaining high quality solutions.

7.5. Large-Scale Machine Learning

In [19], we study large-scale kernel methods for acoustic modeling in speech recognition and compare their performance to deep neural networks (DNNs). We perform experiments on four speech recognition datasets and compare these two types of models on frame-level performance metrics (accuracy, cross-entropy), as well as on recognition metrics (word/character error rate). In order to scale kernel methods to these large datasets, we use the random Fourier feature method. We propose two novel techniques for improving the performance of kernel acoustic models. First, in order to reduce the number of random features required by kernel models, we propose a simple but effective method for feature selection. Second, we present a number of frame-level metrics which correlate very strongly with recognition performance when computed on the heldout set; we take advantage of these correlations by monitoring these metrics during training in order to decide when to stop learning. Additionally, we show that the linear bottleneck method of Sainath et al. improves the performance of our kernel models significantly, in addition to speeding up training and making the models more compact. Together, these three methods dramatically improve the performance of kernel acoustic models, making their performance comparable to DNNs on the tasks we explored.

7.6. Beyond Homophily: Signed networks

In the problem of edge sign prediction, we are given a directed graph (representing a social network), and our task is to predict the binary labels of the edges (i.e., the positive or negative nature of the social relationships). Many successful heuristics for this problem are based on the troll-trust features, estimating at each node the fraction of outgoing and incoming positive/negative edges. In [14], we show that these heuristics can be understood, and rigorously analyzed, as approximators to the Bayes optimal classifier for a simple probabilistic model of the edge labels. We then show that the maximum likelihood estimator for this model approximately corresponds to the predictions of a Label Propagation algorithm run on a transformed version of the original social graph. Extensive experiments on a number of real-world datasets show that this algorithm is competitive against state-of-the-art classifiers in terms of both accuracy and scalability. Finally, we show that troll-trust features can also be used to derive online learning algorithms which have theoretical guarantees even when edges are adversarially labeled.

8. Bilateral Contracts and Grants with Industry

8.1. Product Name Disambiguation

Optimix is a company that provides marketing campaign optimization services and pricing policies for companies. One of the OptiMix tools offers a competitive price comparison. In this collaboration with Magnet, the objective was to use machine learning approaches and natural language processing for product names disambiguation.

8.2. Coreference resolution

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

8.3. 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.

In [17], we presented our preliminary idea for a decentralized and privacy-aware machine learning method for predicting traversal time in the Data Mining with Secure Computing workshop held in conjunction with the 2017 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD-2017).

9. Partnerships and Cooperations

9.1. Regional Initiatives

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

We also 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, FABIO VITALE

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-NFS REM (2016-2020)

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.

9.2.4. 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

ERC-PoC 713626 SOM "Statistical modeling for Optimization Mobility": This project aims at bringing to practice results from the project ERC-StG 240186 MiGraNT in the domain of mobility and mobile devices. In particular, a proof of concept will be made of graph mining approaches to learn predictive models and/or recommendation systems from collections of data distributed over a large number of devices (cars, smartphones, ...) while caring about privacy-friendliness.

9.3.2. Collaborations in European Programs, Except FP7 & H2020

9.3.2.1. Sci-GENERATION (2013-2017)

Program: COST

Project acronym: Sci-GENERATION

Project title: Next Generation of Young Scientist: Towards a Contemporary Spirit of R&I.

Duration: 2013-2017

Coordinator: JAN RAMON is an MC member for belgium and a core group member

Other partners: More information on http://scigeneration.eu/en/participants.html

Abstract: Sci-Generation is a COST targeted network that addresses the challenges faced by next generation of researchers in Europe. We aim to improve the visibility, inclusion and success of excellent young researchers and research teams in European science and policy-making. We study and deliberate how changes in research funding opportunities and career perspectives can facilitate these improvements. We wish to promote new and emergent research topics, methods and management organizations. We are developing recommendations for EU science policy that will foster transformations at national and regional levels to promote scientific excellence and to establish a true European research area. (See http://scigeneration.eu).

9.3.2.2. 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)

Abstract: Effective discourse in any language is characterized by clear relations between sentences and coherent structure. But languages vary in how relations and structure are signaled. While monolingual dictionaries and grammars can characterize the words and sentences of a language and bilingual dictionaries can do the same between languages, there is nothing similar for discourse. For discourse, however, discourse-annotated corpora are becoming available in individual languages. 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. With partners from across Europe, TextLink will unify numerous but scattered linguistic resources on discourse structure. With its resources searchable by form and/or meaning and a source of valuable correspondences, 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 Associate Teams Not Involved in an Inria International Labs

9.4.1.1. RSS

Program: Inria North-European Labs

Project title: Rankings and Similarities in Signed graphs

Duration: late 2015 to late 2017

Partners: Aristides Gionis (Data Mining Group, Aalto University, Finland) and Mark Herbster (Centre for Computational Statistics and Machine Learning, University College London, UK)

Abstract: The project focuses on predictive analysis of networked data represented as signed graphs, where connections can carry either a positive or a negative semantic. The goal of this associate team is to devise novel formal methods and machine learning algorithms towards link classification and link ranking in signed graphs and assess their performance in both theoretical and practical terms.

9.4.1.2. LEGO

Title: LEarning GOod representations for natural language processing

International Partner (Institution - Laboratory - Researcher):

University of Southern California (United States) - Department of Computer Science - 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.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

PETER KLING The objective of the visit of PETER KLING was centered around Learning in Distributed Environments. This initiative contributes to the recent effort of Magnet towards decentralized learning also supported for instance by the Pamela project (Personalized and decentrAlized MachinE Learning under constrAints). Peter Kling's background in distributed computing, combinatorial optimization, online algorithms, and stochastic processes is a good opportunity to investigate new machine learning approaches in this area. In this first of one month, we have started to study Population and Spreading Processes. Two other topics on distributed load balancing and energy-aware algorithms will be the investigated in a second visit in 2018.

VALENTINA ZANTEDESCHI During her one month stay, VALENTINA ZANTEDESCHI has collaborated with AURÉLIEN BELLET and MARC TOMMASI on decentralized learning. A paper on collaborative and decentralized boosting will be submitted in 2018.

ISABEL VALERA visited MAGNET for 3 days to collaborate with AURÉLIEN BELLET on fairness in machine learning.

CLEMENT WEISBECKER visited MAGNET for 1 week to collaborate with AURÉLIEN BELLET on large-scale kernel methods using block low-rank approximations.

WILHELMIINA HAMALAINEN visited MAGNET for 2 weeks to collaborate with JAN RAMON. In particular, they worked on multiple hypothesis tests for regression and discretization problems.

BERT CAPPELLE visited MAGNET for a semester, as part of his "delegation", to collaborate with PASCAL DENIS and MIKAELA KELLER on compositional distributional semantics, and more specifically on the distributional analysis of so-called privative adjectives. A collaborative paper on this work will be submitted in 2018.

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

- R. Babbar (Max Planck Institute): Algorithms for Extreme Multi-Class and Multi-Label Classification
- M. Chehreghani (Xerox Research): Unsupervised Learning over Graphs: Distances, Algorithms, and an Information-Theoretic Model Validation Principle
- G. Boleda (University Pompeu Fabra): Instances and Concepts in Distributional Space
- M. Blondel (NTT): A Regularized Framework for Sparse and Structured Neural Attention
- L. Wehenkel (University of Liège): Probabilistic Reliability Management of the European Electric Power System
- A. Herbelot (University Pompeu Fabra): A Formal Distributional Semantics for Cognitively-Plausible Reference Acts
- H. Ivey-Law (Data61/CSIRO): Private Federated Learning on Vertically Partitioned Data via Entity Resolution and Additively Homomorphic Encryption

9.5.1.1. Internships

Juhi Tandon worked on developing re-ranking parsing models that exploit and compare various tree kernels in the context of semi-supervised graph-based multilingual dependency parsing.

Quentin Tremouille worked on applications of the Hypernode graphs model [39] in the context of (movie) recommendation based on reviews in natural language.

Hippolyte Bourel worked on the application of the decentralized learning algorithms [15] for mobility data.

Rumei Li worked on a Yanakakis style algorithm for computing the effective sample size of a set of dependent training examples.

9.5.2. Visits to International Teams

9.5.2.1. Research Stays Abroad

MATHIEU DEHOUCK visited USC during one month. He worked with pairs of 8 main and auxiliary NLP tasks. More specifically, he looked at transfer learning from low-level tasks (such as part-of-speech tagging, named entity recognition, chunking, word polarity classification) to high-level tasks (e.g., semantic relatedness, textual entailment, sentiment analysis). In contrast to a common belief in the NLP community that transfer learning between these tasks should be possible, we discovered that the widely-used technique in which word representations act as a medium of transfer only leads to limited improvements. These results were presented by Fei Sha at the Inria SiliconValley workshop (BIS'2017), and a paper is in preparation for 2018.

AURÉLIEN BELLET visited École Polytechnique Fédérale de Lausanne (EPFL) during 1 week. He worked with the distributed computing group of Rachid Guerraoui on decentralized and privacy-preserving machine learning, leading to some joint papers [18], [16].

AURÉLIEN BELLET and PASCAL DENIS visited USC during two weeks in December 2017. In collaboration with MELISSA AILEM, recently recruited as a post-doc on the LEGO project, they worked on developing a new algorithm for joint learning of word and image embeddings inspired on the Skip-Gram word2vec model. In addition, they furthered the work initiated with MATHIEU DEHOUCK along with USC colleagues on multi-task learning by proposing a new encoder-decoder model that integrates task and domain embeddings.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. Local Workshops

AURÉLIEN BELLET organized the workshop Decentralized Machine Learning, Optimization and Privacy. More than 50 attendees were present for a series of invited talks on Sep 11-12, 2017. The list of speakers was

- Dan Alistarh (IST Austria / ETH Zurich)
- Borja Balle (Amazon Research)
- Keith Bonawitz (Google Research)
- Hamed Haddadi (QMUL / Imperial College London)
- Stephen Hardy (Data61 / CSIRO)
- Mikael Johansson (KTH)
- Peter Richtárik (KAUST / University of Edinburgh)
- Meilof Veeningen (Philips Research)

MIKAELA KELLER co-organized the local workshop Demi-journée Dating bringing together researchers from the local teams Links, Magnet, Sequel and Sigma for an afternoon of presentations and scientific discussions.

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

AURÉLIEN BELLET served as PC member of NIPS'17, ICML'17, AISTATS'18 and MOD'17.
MARC TOMMASI served as senior PC member of IJCAI'17 and PC member of CAP'17, EGC'17, CRI'17, AAAI'18.
MIKAELA KELLER served as PC member of NIPS'17 and CAP'17.
JAN RAMON served as PC member of AISTATS'17, AISTATS'18, IEEE-BigData'17, BNaic'17, CIKM'17, DS'17, ECML/PKDD'17, ICHI'17, IJCAI'17, ISMIS'17, KDD'17, MLG'17, MOD'17, NIPS'17 and TDLSG-ECMLPKDD'17.
RÉMI GILLERON served as PC member of CAP'17, IJCAI'17, NIPS'17, AISTATS'18, ICLR'18.
PASCAL DENIS served as PC member of ACL-17, CAp'17, EACL'17, EMNLP'17, IJCAI'17, IWCS'17, CORBON'17, and NIPS'17.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

JAN RAMON was member of the editorial board of Machine Learning Journal.
JAN RAMON was member of the editorial board of Data Mining and Knowledge Discovery.

10.1.3.2. Reviewer - Reviewing Activities

JAN RAMON acted as reviewer for project proposals and project follow-up for H2020, COST and SNF (CH).
AURÉLIEN BELLET was reviewer for Machine Learning Journal.

10.1.4. Invited Talks

MARC TOMMASI was invited at the seminar of the machine learning group at the University of Liege.
JAN RAMON gave an invited talk at the TDLSG'17 workshop of ECML/PKDD'17.

AURÉLIEN BELLET gave invited talks at the DALI 2017 Workshop on Fairness and Privacy in Machine Learning, [1] the 2nd Russian-French Workshop in Big Data and Applications, [2] and the Journée Apprentissage et Interactions du GdR IA. [3]
AURÉLIEN BELLET gave an invited lecture at the Pre-doc Summer School on Learning Systems (MPI/ETH). [4]

AURÉLIEN BELLET was invited at the seminars of the Distributed Computing Lab at EPFL, Multispeech (Inria Nancy), Sequel (Inria Lille), SIGMA (Centrale Lille), Proba/Stat (University of Lille), CRIStAL Lille (DaTinG department day).

10.1.5. Scientific Expertise

MARC TOMMASI was member of the ANR CES 23 committee and served in the HCERES committee for the scientific evaluation of a French laboratory.

10.1.6. Research Administration

PASCAL DENIS was member of the Commission Emploi et Recherche (CER) and Commission de Développement Technologique (CDT) at the Inria Lille Research Center.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Licence MIASHS: MARC TOMMASI, Réseaux, 24h, L1, Université Lille 3.

Licence MIASHS: RÉMI GILLERON, Traitement de données, 24h, L1, Université Lille 3.

Licence MIASHS: MATHIEU DEHOUCK et THIBAULT LIÉTARD, Projet informatique de traitement de données en SHS, 20h, L2, Université Lille 3.

Licence MIASHS: MIKAELA KELLER, Codage et représentation de l'information, 24h, L1, Université Lille 3.

Licence SoQ (SHS): FABIEN TORRE, Traitement de contenus textuels, 24h, L3, Université Lille 3.

Licence SoQ (SHS): MIKAELA KELLER, Algorithmique de graphes, 24h, L3, Université Lille 3.

Licence SHS: MIKAELA KELLER, Langages du Web, 24h, L2, Université Lille 3.

Licence SHS: MIKAELA KELLER, Représentation numérique de l'information, 24h, L2-L3, Université Lille 3.

Licence économie gestion: RÉMI GILLERON, Traitement de données et documents, 24h, L1, Université Lille 3.

Licence MARC TOMMASI C2i, Université Lille 3.

Master MIASHS: RÉMI GILLERON et FABIEN TORRE, Web et référencement, 24h, M1, Université Lille 3.

Master MIASHS: GÉRAUD LE FALHER, Web et réseaux, 24h, M1, Université Lille 3.

Master MIASHS: MIKAELA KELLER, Programmation et bases de données, 24h, M1, Université Lille 3.

Master MIASHS: MIKAELA KELLER, Algorithmes fondamentaux de la fouille de données, 60h, M1, Université Lille 3.

Master MIASHS: MARC TOMMASI, Apprentissage et émergence de comportements, 30h, M2, Université Lille 3.

[1] http://dalimeeting.org/dali2017/fairness-and-privacy.html
[2] https://bi.hse.ru/en/rfw/2017/
[3] http://www.gdria.fr/jai17/
[4] http://learning-systems.org/events/summer-school-on-learning-systems-2017

Master LTTAC: FABIEN TORRE, Algorithmique des textes – Javascript, 36h, M1, Université Lille 3.

Master ID: FABIEN TORRE, information structurée, 20h, M2, Université Lille 3.

Master ID: FABIEN TORRE, programmation Web, 20h, M2, Université Lille 3.

Master / Master Spécialisé Big Data: AURÉLIEN BELLET, Advanced Machine Learning, 25.5h, Télécom ParisTech.

Formation continue (Certificat d'Études Spécialisées Data Scientist): AURÉLIEN BELLET, Supervised Learning and Support Vector Machines, 10h, Télécom ParisTech.

Formation continue: AURÉLIEN BELLET, Graph Mining, 3h, Télécom ParisTech pour Allianz.

E-learning

SPOC: MARC TOMMASI, RÉMI GILLERON and ALAIN PREUX: Culture numérique, 5 semesters at the bachelor level, Moodle, Lille 3 university, more than 7000 students.

Pedagogical resources: texts, videos, quiz and exercises available on http://culturenumerique.univ-lille3.fr/, creative commons.

10.2.2. Supervision

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

Postdoc in progress: BO LI, ANR-NFS REM postdoctoral grant, supervised by PASCAL DENIS.

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: MATHIEU DEHOUCK, Graph-based Learning for Multi-lingual and Multi-domain Dependency Parsing, since Oct 2015, PASCAL DENIS and MARC TOMMASI.

PhD in progress: THIBAULT LIÉTARD, Adaptive Graph Learning with Applications to Natural Language Processing, AURÉLIEN BELLET, PASCAL DENIS and RÉMI GILLERON.

PhD in progress: GÉRAUD LE FALHER, Machine Learning in Signed Graphs, Inria Lille – Nord Europe, since Oct. 2014, MARC TOMMASI, FABIO VITALE and CLAUDIO GENTILE (University of Insubria, Italy).

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).

PhD: PAULINE WAUQUIER, Task driven representation learning, Université de Lille, May 29th 2017, MIKAELA KELLER and MARC TOMMASI.

PhD: DAVID CHATEL, Semi-supervised spectral clustering defended, Dec. 7th 2017, MARC TOMMASI and PASCAL DENIS.

10.2.3. Juries

- AURÉLIEN BELLET was member of the PhD committee of ROMAIN BRAULT (Examinateur).
- MARC TOMMASI was member (head) of the recruitment committee for Professor and Assistant Professors in Computer Science at Université Lille 3.
- MARC TOMMASI was member of the habilitation committee of JESSE READ (Rapporteur).
- MIKAELA KELLER was a member of the recruitment committee for the Assistant Professor position in Computer Science at Université Lille 3.
- MIKAELA KELLER was a member of the PhD committee of DAMIEN FOURURE (Examinatrice).
- JAN RAMON was a member of the PhD committee of GIANNIS NIKOLENZOS (Athens, Greece).
- RÉMI GILLERON was a member of the PhD committee of LUDOVIC DOS SANTOS (Rapporteur).

10.3. Popularization

- AURÉLIEN BELLET participated in a discussion on Artificial Intelligence at the Fête de la Science event held at Cité des Sciences (Paris, October 8).
- JAN RAMON presented at the "Pint of Science" series in Lille on the theme of big data and data protection on May 15th.
- JAN RAMON gave a presentation on exploiting traffic data (and the MAGNET work in the MUST project in this domain) in the Inria Meet-up series on December 14th.

11. Bibliography

Major publications by the team in recent years

[1] A. FRENO, M. KELLER, M. TOMMASI. *Fiedler Random Fields: A Large-Scale Spectral Approach to Statistical Network Modeling*, in "Neural Information Processing Systems (NIPS)", Lake Tahoe, United States, Advances in Neural Information Processing Systems, MIT Press, December 2012, vol. 25, https://hal.inria.fr/hal-00750345

[2] O. KUŽELKA, Y. WANG, J. RAMON. *Bounds for Learning from Evolutionary-Related Data in the Realizable Case*, in "International Joint Conference on Artificial Intelligence (IJCAI)", New York, United States, Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 2016, July 2016, https://hal.archives-ouvertes.fr/hal-01422033

[3] E. LASSALLE, P. DENIS. *Improving pairwise coreference models through feature space hierarchy learning*, in "ACL 2013 - Annual meeting of the Association for Computational Linguistics", Sofia, Bulgaria, Association for Computational Linguistics, August 2013, https://hal.inria.fr/hal-00838192

[4] E. LASSALLE, P. DENIS. *Joint Anaphoricity Detection and Coreference Resolution with Constrained Latent Structures*, in "AAAI Conference on Artificial Intelligence (AAAI 2015)", Austin, Texas, United States, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015), January 2015, https://hal.inria.fr/hal-01205189

[5] G. PAPA, S. CLÉMENÇON, A. BELLET. *On Graph Reconstruction via Empirical Risk Minimization: Fast Learning Rates and Scalability*, in "Annual Conference on Neural Information Processing Systems (NIPS 2016)", Barcelone, Spain, December 2016, https://hal.inria.fr/hal-01367546

[6] 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, Paper accepted for publication at ECML/PKDD 2014, https://hal.inria.fr/hal-01017025

Publications of the year

Doctoral Dissertations and Habilitation Theses

[7] D. CHATEL. *Semi-supervised clustering in graphs*, Université de Lille, December 2017, https://hal.inria.fr/tel-01667429

[8] P. WAUQUIER. *Task driven representation learning*, Université Charles de Gaulle - Lille III, May 2017, https://tel.archives-ouvertes.fr/tel-01622153

Articles in International Peer-Reviewed Journals

[9] P. BESSON, N. CARRIÈRE, S. K. BANDT, M. TOMMASI, X. LECLERC, P. DERAMBURE, R. LOPES, L. TYVAERT. *Whole-brain high-resolution structural connectome: inter-subject validation and application to the anatomical segmentation of the striatum*, in "Brain Topography", May 2017, vol. 30, n° 3, pp. 291-302 [*DOI :* 10.1007/s10548-017-0548-0], https://hal.archives-ouvertes.fr/hal-01657960

[10] T. LE VAN, S. NIJSSEN, M. VAN LEEUWEN, L. DE RAEDT. *Semiring Rank Matrix Factorization*, in "IEEE Transactions on Knowledge and Data Engineering", August 2017, vol. 29, n° 8, pp. 1737 - 1750 [*DOI :* 10.1109/TKDE.2017.2688374], https://hal.inria.fr/hal-01666755

[11] C. PELEKIS, J. RAMON, Y. WANG. *Hölder-type inequalities and their applications to concentration and correlation bounds*, in "Indagationes Mathematicae", 2017, vol. 28, n° 1, pp. 170–182 [*DOI :* 10.1016/j.indag.2016.11.017], https://hal.archives-ouvertes.fr/hal-01421953

International Conferences with Proceedings

[12] M. DEHOUCK, P. DENIS. *Delexicalized Word Embeddings for Cross-lingual Dependency Parsing*, in "EACL", Valencia, Spain, EACL 2017, April 2017, vol. 1, pp. 241 - 250 [*DOI :* 10.18653/v1/E17-1023], https://hal.inria.fr/hal-01590639

[13] P. DENIS, L. RALAIVOLA. *Online Learning of Task-specific Word Representations with a Joint Biconvex Passive-Aggressive Algorithm*, in "European Chapter of the Association for Computational Linguistics", Valencia, Spain, April 2017, pp. 775 - 784 [*DOI :* 10.18653/v1/E17-1073], https://hal.inria.fr/hal-01590594

[14] G. LE FALHER, N. CESA-BIANCHI, C. GENTILE, F. VITALE. *On the Troll-Trust Model for Edge Sign Prediction in Social Networks*, in "AISTATS 2017 - 20th International Conference on Artificial Intelligence and Statistics", Fort Lauderdale, United States, April 2017, https://hal.inria.fr/hal-01667039

[15] P. VANHAESEBROUCK, A. BELLET, M. TOMMASI. *Decentralized Collaborative Learning of Personalized Models over Networks*, in "International Conference on Artificial Intelligence and Statistics (AISTATS)", Fort Lauderdale, Florida., United States, April 2017, https://arxiv.org/abs/1610.05202 , https://hal.inria.fr/hal-01533182

Conferences without Proceedings

[16] A. BELLET, R. GUERRAOUI, M. TAZIKI, M. TOMMASI. *Personalized and Private Peer-to-Peer Machine Learning*, in "NIPS 2017 Workshop on Machine Learning on the Phone and other Consumer Devices", Long Beach, United States, 2017, https://hal.inria.fr/hal-01665422

[17] T. LE VAN, A. BELLET, J. RAMON. *Decentralised and Privacy-Aware Learning of Traversal Time Models*, in "ECML PKDD 2017 - workshop DMSC - Data Mining with Secure Computation", Skopje, Macedonia, September 2017, pp. 1-5, https://hal.inria.fr/hal-01666739

Research Reports

[18] A. BELLET, R. GUERRAOUI, M. TAZIKI, M. TOMMASI. *Fast and Differentially Private Algorithms for Decentralized Collaborative Machine Learning*, Inria Lille, 2017, pp. 1-18, https://arxiv.org/abs/1705.08435 , https://hal.inria.fr/hal-01665410

[19] 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*, Inria Lille, 2017, pp. 1-31, https://arxiv.org/abs/1701.03577 , https://hal.inria.fr/hal-01665417

[20] W. ZHENG, A. BELLET, P. GALLINARI. *A Distributed Frank-Wolfe Framework for Learning Low-Rank Matrices with the Trace Norm*, Inria Lille, 2017, https://arxiv.org/abs/1712.07495 , https://hal.inria.fr/hal-01672066

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