IN PARTNERSHIP WITH:
CNRS

Université Charles de Gaulle (Lille 3)

Activity Report 2016

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:

1.2.9. - Social Networks
3. - Data and knowledge
3.1. - Data
3.1.3. - Distributed data
3.1.4. - Uncertain data
3.2.3. - Inference
3.2.4. - Semantic Web
3.3. - Data and knowledge analysis
3.3.1. - On-line analytical processing
3.3.3. - Big data analysis
3.4. - Machine learning and statistics
3.4.1. - Supervised learning
3.4.2. - Unsupervised learning
3.4.4. - Optimization and learning
3.5. - Social networks
3.5.1. - Analysis of large graphs
3.5.2. - Recommendation systems
4.8. - Privacy-enhancing technologies
5.8. - Natural language processing
6.2.6. - Optimization
6.3.1. - Inverse problems
7. - Fundamental Algorithmics
7.2. - Discrete mathematics, combinatorics
7.8. - Information theory
7.9. - Graph theory
7.10. - Network science
7.11. - Performance evaluation
8.1. - Knowledge
8.2. - Machine learning
8.4. - Natural language processing
8.6. - Decision support

Other Research Topics and Application Domains:

1. - Life sciences
1.1.11. - Systems biology
2. - Health
2.2.4. - Infectious diseases, Virology
2.3. - Epidemiology

1. Members

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

Faculty Members
Marc Tommasi [Team leader, Univ. Lille III, Professor, HDR]
Remi Gilleron [Univ. Lille III, Professor, HDR]
Mikaela Keller [Univ. Lille III, Associate Professor]
Fabien Torre [Univ. Lille III, Associate Professor]
Fabio Vitale [Univ. Lille III, Associate Professor]

PhD Students
David Chatel [Univ. Lille I]
Mathieu Dehouck [Univ. Lille I]
Geraud Le Falher [Inria]
Thibault Lietard [Normale Sup Rennes, from Feb 2016]
Pauline Wauquier [Clic and Walk]

Visiting Scientists
Soravit Changpinyo [USC, Oct 2016]
Wilhelmiina Hamalainen [University of Helsinki, Nov 2016]

Administrative Assistant
Julie Jonas [Inria]

Others
Pierre Dellenbach [Univ. Lille III, student intern, from Mar 2016 until Aug 2016]
Paul Vanhaesebrouck [Univ. Lille III, student intern, from Mar 2016 until Jul 2016]

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 [40], [43].

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 [22], [45].

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 [44].

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 [22], [48]. 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 [45], [33]. 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 [35].

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 [34], [30], [32]). 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 [47].

We have already initiated some work on the coreference resolution with graph-based learning, by casting the problem as an instance of spectral clustering [32].

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 [42], face recognition [31], and text categorization [36].

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 ([49], [23], [24]). 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 ([25], [26]) 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 [38]. 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 [46].

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 [37], 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 [28]. 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 bandwith 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 ([29], [39]) 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 [41]. 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. Targeted Applications

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.

5. Highlights of the Year

5.1. Highlights of the Year

- We have been successful in many calls: ERC PoC project SOM (JAN RAMON leader), ANR project GRASP (PASCAL DENIS leader), ANR project PAMELA (MARC TOMMASI is the scientific coordinator), ANR project REM (PASCAL DENIS local leader), ADEME project MUST (JAN RAMON leader), Inria Associate Team LEGO (AURÉLIEN BELLET local leader).

- Scientific advances have been recognized by the community, in top ranked conferences and journals such as ICML, NIPS, JMLR, EMNLP, EACL and IJCAI.

5.1.1. Awards

- CHLOÉ BRAUD, who was supervised by PASCAL DENIS from 2012 to 2015, received the 2016 PhD Award from ATALA, the French NLP association.
- PAUL VANAESEBROUCK, who was supervised par AURÉLIEN BELLET and MARC TOMMASI, has reveived the "Grand Prix du stage de Recherche" from École Polytechnique Paris for his internship in MAGNET (see Section 7.1).

6. New Software and Platforms

6.1. CoRTex

Python library for noun phrase COreference Resolution in natural language TEXts
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.

- Participants: Pascal Denis and David Chatel
- Contact: Pascal Denis
- URL: https://team.inria.fr/magnet/software/

6.2. Magneto

Python toolbox for generating and evaluating vector space representations for Natural Language Processing
FUNCTIONAL DESCRIPTION

Version 1.0 of Magneto contains preprocessing methods for texts in french and english. It includes classical methods for generating vector space representations: count based models, dimensionality reduction based methods and predictive methods (word2vec and Glove). For version 1.0, vector space representations can be evaluated on dedicated evaluation tasks such as similarity and analogy.

- Participants: Pascal Denis, Rémi Gilleron, Mikaela Keller, François Noyer and Nathalie Vauquier
- Contact: Pascal Denis
- URL: https://team.inria.fr/magnet/software/

7. New Results

7.1. Decentralized and Private Learning

In [13], we address the problem of decentralized minimization of pairwise functions of the data points, where these points are distributed over the nodes of a graph defining the communication topology of the network. This general problem finds applications in ranking, distance metric learning and graph inference, among others. We propose new gossip algorithms based on dual averaging which aims at solving such problems both in synchronous and asynchronous settings. The proposed framework is flexible enough to deal with constrained and regularized variants of the optimization problem. Our theoretical analysis reveals that the proposed algorithms preserve the convergence rate of centralized dual averaging up to an additive bias term. We present numerical simulations on Area Under the ROC Curve (AUC) maximization and metric learning problems which illustrate the practical interest of our approach.

In [19], 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. Our algorithm for solving this challenging optimization problem relies on the Alternating Direction Method for Multipliers (ADMM).

In [20], we propose a decentralized protocol for a large set of users to privately compute averages over their joint data, which can later be used to learn more complex models. Our protocol can find a solution of arbitrary accuracy, does not rely on a trusted third party and preserves the privacy of users throughout the execution in both the honest-but-curious and malicious adversary models. 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.

7.2. Natural Language Processing

In [12], we introduce a simple semi-supervised approach to improve implicit discourse relation identification. This approach harnesses large amounts of automatically extracted discourse connectives along with their arguments to construct new distributional word representations. Specifically, we represent words in the space of discourse connectives as a way to directly encode their rhetorical function. Experiments on the Penn Discourse Treebank demonstrate the effectiveness of these task-tailored representations in predicting implicit discourse relations. Our results indeed show that, despite their simplicity, these connective-based representations outperform various off-the-shelf word embeddings, and achieve state-of-the-art performance on this problem.

Along the PhD thesis of THIBAULT LIÉTARD, we are working on learning a similarity between text entities for the task of coreference resolution. Unlike indirect classification criteria often used in the literature, the similarity function naturally operates on pairs of mentions and several relevant objectives can be considered. For instance, we can learn the parameters of the similarity function such that the similarity of a given mention to its closest antecedent coreferent mention is larger than to any closer non-coreferential antecedent candidate. The resulting similarity scores can then be plugged into a greedy clustering procedure, or used to build a weighted graph of mentions to be clustered by spectral algorithms. For the representations of (pairs of) mentions on which the similarity function is learned, we consider both traditional linguistic features as well as external information about the general context of occurrence of the mentions using word embeddings.

Along the PhD thesis of MATHIEU DEHOUCK, we study 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 embeddings, which can be trained on any set of languages and capture features shared accross 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 embeddings (monolingual or multilingual) achieves significant improvements over monolingual baselines. The work is submitted.

7.3. Edge Prediction in Networks

In [18] we address the problem of classifying the links of signed social networks given their full structural topology. 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. 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.

In [16], we address the problem of predicting connections between a set of data points. We focus on the *graph reconstruction* problem, where the prediction rule is obtained by minimizing the average error over all $n(n-1)/2$ possible pairs of the n nodes of a training graph. Our first contribution is to derive learning rates of order $O(\log n/n)$ for this problem, significantly improving upon the slow rates of order $O(1/\sqrt{n})$ established in the seminal work of [27]. Strikingly, these fast rates are universal, in contrast to similar results known for other statistical learning problems (e.g., classification, density level set estimation, ranking, clustering) which require strong assumptions on the distribution of the data. Motivated by applications to large graphs, our second contribution deals with the computational complexity of graph reconstruction. Specifically, we investigate to which extent the learning rates can be preserved when replacing the empirical reconstruction risk by a computationally cheaper Monte-Carlo version, obtained by sampling with replacement $B \ll n^2$ pairs of nodes. Finally, we illustrate our theoretical results by numerical experiments on synthetic and real graphs.

7.4. Mining Geotagged Social Data

Data generated on location-based social networks provide rich information on the whereabouts of urban dwellers. Specifically, such data reveal who spends time where, when, and on what type of activity (e.g., shopping at a mall, or dining at a restaurant). That information can, in turn, be used to describe city regions in terms of activity that takes place therein. For example, the data might reveal that citizens visit one region mainly for shopping in the morning, while another for dining in the evening. Furthermore, once such a description is available, one can ask more elaborate questions. For example, one might ask what features distinguish one region from another – some regions might be different in terms of the type of venues they host and others in terms of the visitors they attract. As another example, one might ask which regions are similar across cities. In [11], we present a method to answer such questions using publicly shared Foursquare data. Our analysis makes use of a probabilistic model, the features of which include the exact location of activity, the users who participate in the activity, as well as the time of the day and day of week the activity takes place. Compared to previous approaches to similar tasks, our probabilistic modeling approach allows us to make minimal assumptions about the data – which relieves us from having to set arbitrary parameters in our analysis (e.g., regarding the granularity of discovered regions or the importance of different features). We demonstrate how the model learned with our method can be used to identify the most likely and distinctive features of a geographical area, quantify the importance features used in the model, and discover similar regions across different cities. Finally, we perform an empirical comparison with previous work and discuss insights obtained through our findings. Our results were also presented through an interactive demo at the 25th World Wide Web Conference [21].

7.5. Learning from Non-iid Data

In [14] we deal with the generalization ability of classifiers trained from non-iid evolutionary-related data in which all training and testing examples correspond to leaves of a phylogenetic tree. For the realizable case, we prove PAC-type upper and lower bounds based on symmetries and matchings in such trees.

In [9], we studied learning problems where the performance criterion consists of an average over tuples (e.g., pairs or triplets) of observations rather than over individual observations, as in many learning problems involving networked data (e.g., link prediction), but also in metric learning and ranking. In this setting, the empirical risk to be optimized takes the form of a U-statistic, and its terms are highly dependent and thus violate the classic i.i.d. assumption. From a computational perspective, the calculation of such statistics is highly expensive even for a moderate sample size n, as it requires averaging $O(n^d)$ terms. We show that, strikingly, such empirical risks can be replaced by drastically computationally simpler Monte-Carlo estimates based on $O(n)$ terms only, usually referred to as incomplete U-statistics, without damaging the $O(1/\sqrt{n})$ learning rate of Empirical Risk Minimization (ERM) procedures. For this purpose, we establish uniform deviation results describing the error made when approximating a U-process by its incomplete version under appropriate complexity assumptions. Extensions to model selection, fast rate situations and various sampling techniques are also considered , as well as an application to stochastic gradient descent for ERM. Finally, numerical examples are displayed in order to provide strong empirical evidence that the approach we promote largely surpasses more naive subsampling techniques.

7.6. Adaptive Graph Construction

The efficiency of graph-based semi-supervised algorithms depends on the graph of instances on which they are applied. The instances are often in a vectorial form before a graph linking them is built. The construction of the graph relies on a metric over the vectorial space that help define the weight of the connection between entities. The classic choice for this metric is usually a distance measure or a similarity measure based on the euclidean norm. We claim that in some cases the euclidean norm on the initial vectorial space might not be the more appropriate to solve the task efficiently. In the work [17], we propose an algorithm that aims at learning the most appropriate vectorial representation for building a graph on which the task at hand is solved efficiently. In addition to experimental results showing the interest of such an approach, we define initial conditions under which the graph-based classification is ensured to perform optimally.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. Cifre Clic and Walk (2013-2016)

Participants: MIKAELA KELLER [correspondent], PAULINE WAUQUIER, MARC TOMMASI.

We have a one to one cooperation with the CLIC AND WALK company that makes marketing surveys by consumers (called clicwalkers). The goal of the company is to understand the community of clicwalkers (40 thousands in one year) and its evolution with two objectives: the first one is to optimize the attribution of surveys to clicwalkers, and the second is to expand company's market to foreign countries. Social data can be obtained from social networks (G+, Facebook, ...) but there is no explicit network to describe the clicwalkers community. But users activity in answering surveys as well as server logs can provide traces of information diffusion, geolocalisation data, temporal data, sponsorship, etc. We study the problem of adaptive graph construction from the clicwalkers network. Node (users) classification and clustering algorithms are applied. For the problem of survey recommendations, the problem of teams constitution in a bipartite graph of users and surveys is studied. Random graph modeling and generative models of random graphs will be one step towards the prediction of the evolution of clicwalkers community.

8.1.2. ADEME

ADEME project MUST: Méthodologie d'exploitation des donnees d'usage des vehicules et d'identification de nouveaux Services pour les usagers et les territoires. JAN RAMON is the local PI at Inria of this project.

9. Partnerships and Cooperations

9.1. Regional Initiatives

Participation to the *Data Advanced data science and technologies* project (CPER Data). This project, leaded by DAVID SIMPLOT-RYL, 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.

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 organisations. 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.

9.3.2.3. STAC (2011-2016)

Program: ERC Advanced Grant

Project acronym: STAC

Project title: Strategic conversation

Duration: Sep. 2011 - Aug. 2016

Coordinator: Nicholas Asher, CNRS, Université Paul Sabatier, IRIT (France)

Other partners: School of Informatics, Edinburgh University; Heriot Watt University, Edinburgh; Inria (PASCAL DENIS)

Abstract: STAC is a five year interdisciplinary project that aims to develop a new, formal and robust model of conversation, drawing from ideas in linguistics, philosophy, computer science and economics. The project brings a state of the art, linguistic theory of discourse interpretation together with a sophisticated view of agent interaction and strategic decision making, taking advantage of work on game theory.

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 California, Los Angeles (United States) - TEDS: Research group Theoretical and Empirical Data Science - Fei Sha

Start year: 2016

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

Abstract: 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

We invited Soravit Changpinyo (University of Southern California) in October, collaborating with MATHIEU DEHOUCK, PASCAL DENIS and AURÉLIEN BELLET on multi-task learning and transfer of word embeddings.

JAN RAMON collaborated with WILHELMIINA HAMALAINEN, who visited the magnet lab for 2 weeks. In particular, they worked on multiple hypothesis tests for regression and discretization problems.

MARK HERBSTER from University College London was invited for one week in January and collaborated with FABIO VITALE and MARC TOMMASI on machine learning and similarity prediction in graphs.

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

- TIM VANDERCRUYS (Toulouse): "Modeling Meaning with Latent Factorization Models" (April)
- SORAVIT CHANGPINYO (University of Southern California): "Synthesized Classifiers for Zero-Shot Learning" (October)
- THOMAS KIPF (University of Amsterdam): "Deep Learning on Graphs with Graph Convolutional Networks" (December)

9.5.1.1. Local Workshops

- FABIO VITALE organized the workshop Graph-based Learning and Graph Mining.
- PASCAL DENIS organized the Workshop on Argumentation Mining.

9.5.2. Visits to International Teams

In March, April and May FABIO VITALE visited the Department of Computer Science of the University of Milan, collaborating with Prof. NICOLÒ CESA-BIANCHI and Prof. CLAUDIO GENTILE.

In July, AURÉLIEN BELLET and PASCAL DENIS visited the Department of Computer Science of the University of California (Los Angeles), collaborating with Prof FEI SHA.

In September, MATHIEU DEHOUCK visited the Department of Computer Science of the University of California (Los Angeles), collaborating with Prof FEI SHA.

Since September, FABIO VITALE is working at the department of computer science of Aalto University, Helsinki (Finland), in the DMG group (http://research.ics.aalto.fi/dmg/index.shtml) led by Prof. ARISTIDES GIONIS.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. Member of the Organizing Committees

AURÉLIEN BELLET co-organized the workshop Private Multi-Party Machine Learning @ NIPS 2016. [1]

10.1.2. Scientific Events Selection

10.1.2.1. Chair of Conference Program Committees

PASCAL DENIS served as co-chair of the Polaris Colloquium, a monthly Guest Lecture series in Computer Science and Signal Processing, co-sponsored by Inria Lille–Nord Europe and University of Lille CRIStAL Lab.

[1] https://pmpml.github.io/PMPML16/

10.1.2.2. Member of the Conference Program Committees

AURÉLIEN BELLET served as PC member for IJCAI 2016, ICML 2016, NIPS 2016 and AISTATS 2017.

PASCAL DENIS served as Senior PC member for IJCAI 2016. He was PC member for ACL 2016, AAAI 2016, CAp 2016, EMNLP 2016, and NAACL 2016.

RÉMI GILLERON served as PC member for NIPS 2016 and AISTATS 2017.

JAN RAMON served as PC member for AISTATS 2016 and 2017, BAI workshop @ IJCAI 2017, ieee big data 2016, BNAIC 2016, DS 2016, ECAI 2016, ECML-PKDD 2016, IEEE ICDM 2016, ICHI 2016, IJCAI 2016, ILP 2016, ISMIS 2017, KDD 2016, MLG 2016, NIPS 2016, PMPML workshop @ NIPS 2016, SSDM workshop @ ECML-PKDD 2016.

MARC TOMMASI served as PC member for NIPS 2016, ICML 2016, IJCAI 2016, EGC 2017.

FABIEN TORRE served as PC member for EGC 2017, workshop CluCo 2017, AAFD & SFC 2016, CNIA 2016

FABIO VITALE served as PC member for AISTATS 2017 and ECMLPKDD 2016, NIPS 2016.

10.1.2.3. Reviewer

GÉRAUD LE FALHER was reviewer for ECMLPKDD 2016, NIPS 2016 and IJCAI 2016. MIKAELA KELLER was a reviewer for ICLR 2016 and ICML 2016.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

JAN RAMON is member of the editorial boards of data mining and knowledge discovery, machine learning and guest editorial board of ECML-PKDD 2016.

10.1.3.2. Reviewer - Reviewing Activities

AURÉLIEN BELLET was reviewer for IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) and IEEE Transactions on Cybernetics (TCYB).

JAN RAMON: data mining and knowledge discovery (dami) 5; knowledge and information systems (kais): 4; machine learning :3; neurocomputing: 2; artificial intelligence: 2; plos one : 1

FABIO VITALE was reviewer for the journal EJOR (European Journal of Operational Research - Elsevier), and the journal Internet Mathematics.

GÉRAUD LE FALHER was reviewer for the International Journal of E-Planning Research.

10.1.4. Invited Talks

AURÉLIEN BELLET was invited the Learning, Privacy, and Mobile Data Workshop at Google Research Seattle, [2] the Statistical Machine Learning (SMILE) seminar in Paris, [3] the Workshop on Distributed Machine Learning (Télécom ParisTech), [4] and at the company EURA NOVA. [5]

PASCAL DENIS was invited to the STL seminar, Université Lille 3 (February 2016).

10.1.5. Scientific Expertise

PASCAL DENIS was reviewer for Flanders Research Foundation (FWO, Belgium) and the Brussels Institute for Research and Innovation (Innoviris).

JAN RAMON was reviewer of H2020 projects.

10.1.6. Research Administration

FABIEN TORRE is in the board of the national evaluation committee for teaching and research in computer science (CNU 27)

[2] https://sites.google.com/site/learningprivacymobiledata/
[3] https://sites.google.com/site/smileinparis/
[4] http://machinelearningforbigdata.telecom-paristech.fr/fr/article/workshop-friday-novembre-25-distributed-machine-learning
[5] http://euranova.eu

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Licence MIASHS: FABIO VITALE, Introduction à l'algorithmique, 66h, L2, Université Lille 3.

Licence MIASHS: MARC TOMMASI, Réseaux, 28h, L1, Université Lille 3.

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

Licence MIASHS: GÉRAUD LE FALHER, Traitement de données, 36h, 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): RÉMI GILLERON, Algorithmique de graphes, 24h, L3, Université Lille 3.

Licence SHS: MIKAELA KELLER, Langages du Web, 24h, L3, Université Lille 3.

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

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

Licence MARC TOMMASI, MIKAELA KELLER C2i, université Lille 3.

Master MOCAD: PASCAL DENIS co-taught the Machine Learning and Decision under Uncertainty class, 37,5h, Université Lille 1.

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 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, quizz and exercices available on http://culturenumerique.univ-lille3.fr/, creative commons.

10.2.2. Supervision

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: DAVID CHATEL, Semi-supervised spectral clustering since Sep 2012, MARC TOMMASI and PASCAL DENIS.

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: PAULINE WAUQUIER, Recommendation in Information Networks, since Dec 2013, MIKAELA KELLER and MARC TOMMASI.

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

Master: THIBAULT LIÉTARD, Metric learning for Graph-based coreference resolution, ENS Rennes, co-supervised by AURÉLIEN BELLET and PASCAL DENIS.

Master: PAUL VANAESEBROUCK, Decentralized Machine Learning on Graphs, Ecole Polytechnique, co-supervised by AURÉLIEN BELLET and MARC TOMMASI.

Master: PIERRE DELLENBACH, Private learning of heavily distributed data, Ecole Polytechnique, co-supervised by AURÉLIEN BELLET and JAN RAMON.

Master: ROBIN VOGEL, Learning to Rank Rare Instances, ENSAE, co-supervised by AURÉLIEN BELLET, STÉPHAN CLÉMENÇON et ANNE SABOURIN (Télécom ParisTech), et STÉPHANE GENTRIC (Morpho).

10.2.3. Juries

- AURÉLIEN BELLET was member of the recruitment committee for MdC in Computer Science at Télécom Saint-Etienne.
- PASCAL DENIS et MARC TOMMASI were members of the recruitment committee for MdC in Computer Science at Université Lille 3.
- PASCAL DENIS was member of the Commission Emploi Rechercher (CER) at Inria Lille – Nord Europe.
- RÉMI GILLERON was member of the PhD committees of HUGO LOUCHE (Examinateur) and TOM SEBASTIAN (Président).
- MARC TOMMASI was member of the habilitation committee of ALBERT BIFFET (Rapporteur).
- MARC TOMMASI was member of the PhD committees of GUILLAUME RABUSSEAU (Rapporteur), RAPHAËL PUGET (Rapporteur), MICHAËL PERROT (Examinateur), HADRIEN GLAUDE (Examinateur).
- MARC TOMMASI was head of the jury for the recruitment committee of Junior Research Scientists (CR1/CR2) at Inria Lille.

10.3. Popularization

MARC TOMMASI presented the MAGNET team at the EuraTechnologies ICT innovation ecosystem (https://www.inria.fr/centre/lille/agenda/r-dv-du-plateau-inria-apprentissage-automatique-et-reseaux-d-information).

AURÉLIEN BELLET presented some of his work at the popularization seminar “30 minutes de sciences” and at an Assemblée Générale of Inria Lille - Nord Europe.

THIBAULT LIÉTARD has participated to the “chercheur itinérant” initiative (https://www.inria.fr/centre/lille/recherche/sciences-pour-tous2/mediation/chercheurs-itinerants-au-lycee-2016).

RÉMI GILLERON has participated at the forum “F O O R <: Forum Ouvert Oeuvres et Recherches” where he presented the research work done with PASCAL DENIS for the artwork “This is Major Tom to Ground Control”

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] C. PELEKIS, J. RAMON, Y. WANG. *Hölder-type inequalities and their applications to concentration and correlation bounds*, in "Indagationes Mathematicae", 2016

[7] 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

Articles in International Peer-Reviewed Journals

[8] A. BELLET, J. F. BERNABEU, A. HABRARD, M. SEBBAN. *Learning Discriminative Tree Edit Similarities for Linear Classification - Application to Melody Recognition*, in "Neurocomputing", 2016, vol. 214, pp. 155-161 [*DOI :* 10.1016/J.NEUCOM.2016.06.006], https://hal.archives-ouvertes.fr/hal-01330492

[9] S. CLÉMENÇON, I. COLIN, A. BELLET. *Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics*, in "Journal of Machine Learning Research (JMLR)", 2016, vol. 17, n^{o} 76, pp. 1-36, https://hal.inria.fr/hal-01327662

[10] E. MAES, P. KELCHTERMANS, W. BITTREMIEUX, K. DE GRAVE, S. DEGROEVE, J. HOOYBERGHS, I. MERTENS, G. BAGGERMAN, J. RAMON, K. LAUKENS, L. MARTENS, D. VALKENBORG. *Designing biomedical proteomics experiments: state-of-the-art and future perspectives*, in "Expert Review of Proteomics", 2016 [*DOI :* 10.1586/14789450.2016.1172967], https://hal.inria.fr/hal-01431414

[11] E. ÇELIKTEN, G. C. LE FALHER, M. C. MATHIOUDAKIS. *Modeling Urban Behavior by Mining Geotagged Social Data*, in "IEEE Transactions on Big Data", December 2016, 14 p. [*DOI :* 10.1109/TBDATA.2016.2628398], https://hal.inria.fr/hal-01406676

International Conferences with Proceedings

[12] C. BRAUD, P. DENIS. *Learning Connective-based Word Representations for Implicit Discourse Relation Identification*, in "Empirical Methods on Natural Language Processing", Austin, United States, November 2016, https://hal.inria.fr/hal-01397318

[13] I. COLIN, A. BELLET, J. SALMON, S. CLÉMENÇON. *Gossip Dual Averaging for Decentralized Optimization of Pairwise Functions*, in "International Conference on Machine Learning (ICML 2016)", New York, United States, June 2016, https://hal.inria.fr/hal-01329315

[14] 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

[15] Z. LU, D. GUO, A. B. GARAKANI, K. LIU, A. MAY, A. BELLET, L. FAN, M. COLLINS, B. KINGSBURY, M. PICHENY, F. SHA. *A Comparison Between Deep Neural Nets and Kernel Acoustic Models for Speech Recognition*, in "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)", Shanghai, China, March 2016, https://hal.inria.fr/hal-01329772

[16] 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

Conferences without Proceedings

[17] P. WAUQUIER, M. KELLER. *A Metric Learning Approach for Graph-Based Label Propagation*, in "Workshop track of ICLR 2016", San Juan, Puerto Rico, May 2016, https://hal.inria.fr/hal-01427287

Research Reports

[18] G. LE FALHER, N. CESA-BIANCHI, C. GENTILE, F. VITALE. *On the Troll-Trust Model for Edge Sign Prediction in Social Networks*, Inria Lille, 2016, https://hal.inria.fr/hal-01425137

[19] P. VANHAESEBROUCK, A. BELLET, M. TOMMASI. *Decentralized Collaborative Learning of Personalized Models over Networks*, Inria Lille, October 2016, https://hal.inria.fr/hal-01383544

Other Publications

[20] P. DELLENBACH, J. RAMON, A. BELLET. *A Decentralized and Robust Protocol for Private Averaging over Highly Distributed Data*, December 2016, NIPS 2016 workshop on Private Multi-Party Machine Learning, Poster, https://hal.inria.fr/hal-01384148

[21] E. ÇELIKTEN, G. LE FALHER, M. MATHIOUDAKIS. *"What Is the City but the People?" Exploring Urban Activity Using Social Web Traces*, 25th World Wide Web Conference, Demo Track, April 2016, 25th World Wide Web Conference, Demo Track, Poster [*DOI :* 10.1145/2872518.2901922], https://hal.inria.fr/hal-01295344

References in notes

[22] 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

[23] 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

[24] 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

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

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

[27] G. BIAU, K. BLEAKLEY. *Statistical Inference on Graphs*, in "Statistics & Decisions", 2006, vol. 24, pp. 209–232

[28] 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

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

[30] 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

[31] 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

[32] 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

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

[34] 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

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

[36] 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

[37] 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

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

[39] 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

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

[41] 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

[42] 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

[43] 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

[44] 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

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

[46] 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

[47] 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

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

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