Activity Report 2013

Team MAGNET

Machine Learning in Information Networks

RESEARCH CENTER
Lille - Nord Europe

THEME
Data and Knowledge Representation and Processing

Table of contents

Team MAGNET

Keywords: Machine Learning, Graphs, Networks, Natural Language

Creation of the Team: 2013 January 01.

1. Members

Research Scientists

Gemma Casas Garriga [Researcher, Inria, from Jan 2013 until Feb 2013]
Pascal Denis [Researcher, Inria, since Jan 2013]

Faculty Members

Marc Tommasi [Team leader, Professor, Univ Lille III, since Jan 2013, HdR]
Rémi Gilleron [Professor, Univ Lille III, since Jan 2013, HdR]
Mikaela Keller [Associate Professor, Univ Lille III, since Jan 2013]
Fabien Torre [Associate Professor, Univ Lille III, since Jan 2013]
Fabio Vitale [Associate Professor, Univ Lille III, since Sep 2013]

Engineer

Guillaume Bagan [Inria, from Jan 2013 until May 2013]

PhD Students

Pauline Wauquier [Cifre Clic and Walk, since Dec 2013]
David Chatel [Conseil Régional du Nord-Pas de Calais, Inria, since Jan 2013]
Jean Decoster [Univ Lille I, since Jan 2013]
Seyed Ahmad Hosseini [Univ Lille I, Univ Lille III, Inria, since Sep 2013]
Thomas Ricatte [Cifre SAP, since Jan 2013]

Post-Doctoral Fellow

Antonino Freno [ANR LAMPADA project, Inria, from Jan 2013 until May 2013]

Administrative Assistant

Julie Jonas [Inria, since Jan 2013]

2. Overall Objectives

2.1. Presentation

MAGNET is a new 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. We will then define 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. Envisioned applications include browsing, monitoring and recommender systems, and more broadly information extraction in information networks. Application domains cover social networks for cultural data and e-commerce, and biomedical informatics.

2.2. Highlights of the Year

As first highlight, we are happy to report that our paper "Fiedler Random Fields: A Large-Scale Spectral Approach to Statistical Network Modeling " has been accepted for publication at *Journal of Machine Learning Research*, the top journal in the field of machine learning. This paper's contributions are twofold. First, we introduce the Fiedler delta statistic, based on the Laplacian spectrum of graphs, which allows to dispense with any parametric assumption concerning the modeled network properties. Second, we use the defined statistic to develop the Fiedler random field model, which allows for efficient estimation of edge distributions over large-scale random networks. After analyzing the dependence structure involved in Fiedler random fields, we estimate them over several real-world networks, showing that they achieve a much higher modeling accuracy than other well-known statistical approaches.
The second highlight of the year is the publication of our paper " Improving pairwise coreference models through feature space hierarchy learning" at the annual *Meeting of the Association for Computational Linguistics (ACL 2013)*, the premier conference in the field of Natural Language Processing. This paper proposes a new method for significantly improving the performance of pairwise coreference models. Given a set of indicators, our method learns how to best separate types of mention pairs into equivalence classes for which we construct distinct classification models. In effect, our approach finds an optimal feature space (derived from a base feature set and indicator set) for discriminating coreferential mention pairs. Although our approach explores a very large space of possible feature spaces, it remains tractable by exploiting the structure of the hierarchies built from the indicators. Our experiments on the CoNLL-2012 Shared Task English datasets (gold mentions) indicate that our method is robust relative to different clustering strategies and evaluation metrics, showing large and consistent improvements over a single pairwise model using the same base features.

3. Research Program

3.1. Introduction

The main objective of MAGNET is to develop original machine learning methods for networked data. We consider information networks in which the data are vectorial data and texts. We model such information 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 learning algorithms to build applications like browsing, monitoring and recommender systems, and more broadly information extraction in information networks. Hence, we will investigate new learning algorithms for node clustering and node classification, link classification and link prediction. Also, we will search for the best hidden graph structure to be generated for solving a given learning task. We will base our research 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 and randomization can be used in new machine learning algorithms. Also, active machine learning algorithms for graphs will be investigated.

On the first hand we want to design machine learning algorithms on graphs to solve problems in networks of texts and documents in natural language. The main originality of this research is to consider and take advantage of the setting of networked data exploiting the relationships between different data entities and, overall, the graph topology. On the second hand, in a concomitant way, we want to develop prediction models for graph-like data. This includes prediction, ranking and classification of links and nodes in an on-line or batch setting. The two objectives are intertwined, enrich each other and raise important scientific questions we want to focus on. Our research proposal is organized according to the following questions:

1. How to go beyond vectorial classification models in natural language oriented tasks?
2. How to adaptively build graphs with respect to the given tasks? How to create network from observations of information diffusion processes?

3. How to design methods able to achieve very good predictive accuracy without giving up on scalability?
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. Thus, documents on the web are linked through hyperlinks, forum posts and emails are organized in threads, tweets can be retweeted, etc. Additional connections can be made through users connections (co-authorship, friendship, follower, etc.). Interestingly, NLP research has been rather slow in coming to terms with this situation, and most work 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 [28], [30].

A second motivation comes from the fact that most (if not all) NLP problems can be naturally conceived as graph problems. Thus, NL 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 in principle 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, or at least complement, to structured prediction methods (such as CRFs or structured SVMs) commonly used within NLP. While structured output approaches are able to model local dependencies (e.g., between neighboring words or sentences), they cannot efficiently capture long distance dependencies, like forcing a particular n-gram to receive the same labeling in different sentences or documents for instance. On the other hand, graph-based models provide a natural way to capture global properties of the data through the exploitation of walks and neighborhood in graphs. 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 [10], [32].

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 performances for several NL 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 [31].

Part of the challenge in this work 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, frienship, or thread membership. Unfortunately, modeling NL 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. Of course, there are various well-known ways to obtain similarity measures between text contents (and its associated vectorial data), and graphs can be easily constructed from those combined with some sparsification method. But we would like our similarity to be tailored to the task objective. An additional problem with many NLP problems is that features typically live in different types of spaces (e.g., binary, discrete, continuous). A preliminary discussion of the issue of optimal graph construction for semi-supervised learning in NLP is given in [10], [26]. We identify the issue of adaptative 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 structure prediction problems, allowing to capture (some of the) output dependencies. Structure prediction can be viewed as (set of) link prediction with global loss or dependencies, which means that graph-based learning methods can handle (at least, approximately) output prediction dependencies, and they can in principle capture additional more global dependencies given the right graph structure. 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 based regularization and graph propagation methods. Within such approaches, labels are typically binary or they correspond to 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 [32], [18]. Another related question that we will address in this context is how does one learn with *partial labels* (like partially specificied tag sequence or tree) and use the graph structure to complete the output structure. This last question is very relevant to NL 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 [19].

The NL tasks we will mostly focus on are coreference resolution and entity linking, temporal structure prediction, and discourse parsing. These tasks will be envisonned 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 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. 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 [34].

We have already initiated some work on the coreference resolution problem in the context of ML graph-based approaches. We cast this problem as a spectral clustering problem. Given than features can be numerical or nominal, the definition of a good similarity measure between entities is not straightforward. As a first solution, we consider only numerical attributes to build a k-nn graph of mentions so that graph clustering methods can be applied. Nominal attributes and relations are introduced by means of soft constraints on this clustering. Constraints can have various forms and have the ability of going beyond homophily assumptions, taking into account for instance dissimilarity relationships. From this setting we derive new graph-based learning methods. We propose to study the modification of graph clustering and spectral embeddings to satisfy certain constraints induced by several types of supervision: (i) nodes belong to the same group or to different groups, and (ii) some groups are fully known while others have to be discovered. This semi-supervised graph clustering problem is studied in a batch and transductive setting. But interesting extensions can be investigated in an online and active setting.

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, which makes it possible to infer information related

to each data sample even exploiting the graph topology solely. In fact, a widespread approach to the solution of several classification problems is representing 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 ([29]), face recognition ([17]), and text categorization ([22]).

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 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 performance ([35], [11], [12]). 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. One is the question of choosing the best similarity measure given the objective learning task. This question is related to the question of similarity learning ([13]) 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 adaptative 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 [24]. 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. We have started to work on combining graphs in a semi supervised setting for node classification problems along the PhD thesis of T. Ricatte. Future work include combination geared by semi-supervision on link prediction tasks. This can be studied in an active learning setting. But one important issue is to design scalable approaches, thus to exploit locality given by the network. Doing this we address another objective to build non uniformly parameterized combinations.

3.4. Prediction on Graphs and Scalability

As stated in the previous sections, graphs as complex objects provides 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 recover 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 labelling 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-based 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 smooth labeling function corresponding to an harmonic function on both manifolds in input and output. We also plan to extend our results on spectral clustering with must-link and cannot-link constraints in two directions. We have proposed a batch method with an optimization problem based on an adaptive spectral embedding with respects to constraints. We want to extend this approach to an on-line and active setting where a flow of graphs (each one is a document) is given as input. In the case of large graphs, we also consider the case where partial supervision consists in the knowledge of few clusters.

Scalability is one of the main issue 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 computation time scales quadratically, or slower, in the number of considered data objects (usually nodes or vertices, depending on the given 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 [33]. This approach leaves us with the problem of choosing a good spanning tree, taking into account that the setting could be adversarial (e.g, in the online case the presentation and the assignment of the labels are both arbitrary). A suitable use of the randomization power becomes therefore remarkably significant. Moreover, it is interesting to observe that running a prediction algorithm on a sparsified version of the input dataset allows the parallelization of prediction tasks. In fact, given a prediction task for a networked dataset, in a preliminary phase one could run a randomized graph sparsification method in parallel on different machines. For example, in the case of the spanning tree use, one could then draw several spanning trees at the same time, each on a different computer. This way it is possible to simultaneously run different prediction experiments on the same task and aggregating the obtained results at the end, with several methods (e.g. simply by majority vote) in order to increase the robustness and accuracy predictions.

At the level of the 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 [23], 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 [14]. 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.

We intend to develop new learning models for link prediction problems. We have already proposed a conditional model in [21] with statistics based on Fiedler values computed on small subgraphs. We will investigate the use of such a conditional model for link prediction. We will also extend the conditional probabilistic models to the case of graphs with textual and vectorial data by defining joint conditional models. Indeed, an important challenge for information networks is to introduce node contents in link ranking and link prediction methods that usually rely solely on the graph structure. A first step in this direction was already proposed in [20] where we learn a mapping of node content to a new representation constrained by the existing link structure and applied it for link recommendation. This approach opens a different view on recommendation by means of link ranking problems for which we think that non parametric approaches should be fruitful.

Regarding link classification problems, we plan to devise a whole family of active learning strategies, which could be based on spanning trees or sparse input subgraphs, that exploit randomization and the structure of the graph in order to offset the adversarial label assignment. We expect these active strategies to exhibit good accuracies with a remarkably small number of queried edges, where passive learning methods typically break down. The theoretical findings can be supported by experiments run on both synthetic and real-world (Slashdot, Epinions, Wikipedia, and others) datasets.

We are interested in studying generative models for graph labeling, exploiting the results obtained in p-stochastic model for link classification (investigated in [16]) and statistical model for node label assignment which can be related to tree-structured Markov random fields [25].

In developing our algorithms, we focus on providing theoretical guarantees on prediction accuracy and, at the same time, on computational efficiency. The development of methods that simultaneously guarantee optimal accuracy and computational efficiency is a very challenging goal. In fact, the accuracy of most methods in the literature is not rigorously analyzed from a theoretical point of view. Likewise, tight time and space complexity bounds are not generally provided. This contrasts with the need to manage extremely large relational datasets like, e.g., snapshots of the World Wide Web.

3.5. Beyond Homophilic Relationships

In many cases, the algorithms devised 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 ([15], [27]) 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 interests is one of the most significant reasons 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. Concrete examples are provided by certain types of online social networks. Users of Slashdot can tag other users as friends or foes. Similarly, users of Epinions can give positive or negative ratings not only to products but also to other users. Even in the social network of Wikipedia administrators, votes cast by an admin in favor or against the promotion of another admin can be viewed as positive or negative links. More examples of signed links are found in other domains, such as the excitatory or inhibitory interactions between genes or gene products in biological networks.

Although the introduction of signs on graph edges appears like a small change from standard weighted graphs, the resulting mathematical object, called signed graph, has an unexpectedly rich additional complexity. For example, the spectral properties of signed graphs, which essentially all sophisticated node classification algorithms rely on, are different and less known than those of their unsigned counterparts. Signed graphs naturally lead to a specific inference problem that we have discussed in previous sections: link classification. This is the problem of predicting the sign 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 relationship between individuals.

Another way to go beyond homophily between entities will be studied using our recent model of hypergraphs with bipartite hyperedges [8]. 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 permits to weight the participation of each node to the relationship modeled by bipartite hyperedges accordingly to an equilibrium condition. This is exactly that equilibrium condition that provides a competition between nodes in hyperedges and allows interesting modeling properties that go beyond homophily and similarity over nodes. (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 in which 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.

To conclude, we plan to go beyond the homophilic bias from the algorithmic as well as from the modeling point of view. We will consider new kind of modeling and learning biases provided by graphs with negative weights (signed graphs) and hypergraphs. We will study their spectral properties, smoothness measures of (node or edge) labeling. Sampling and walking also need to be reconsidered. From the machine learning perspective, we will study edge and node labeling in batch and online settings. In connection with our main targeted applications, we will mainly consider unsupervised and semi-supervised situations. We think that allowing negative weights and advanced relationships on nodes will also lead to space efficient representations of graphs.

4. Application Domains

4.1. Overview

Our main targeted applications are browsing, monitoring and mining in information networks. Such discovered structures would also be beneficial to predicting links between users and texts which is at the core of recommender systems. All 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. Software and Platforms

5.1. CoRTex

Participants: Pascal Denis [correspondent], David Chatel.

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 developped in my own research, such as [3]. 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. This project is hosted on Inria gforge: https://gforge.inria.fr/projects/cortex/.

5.2. JProGraM

Participant: Antonino Freno [correspondent].

JProGraM is a GPL-licensed Java library for machine learning and statistical analysis over graphs and through graphs. Supported models for vectorial data include e.g. Bayesian networks, Markov random fields, Gaussian mixtures, kernel density estimators, and neural networks, whereas random graph tools include small-world networks, preferential-attachment, exponential random graphs, and spectral models (as well as subgraph sampling algorithms). One strong point of the library is the extensive support for continuous random variables. JProGraM integrates implementations for the recent results in [20] and [21]. For more information, see the associated webpage at http://researchers.lille.inria.fr/~freno/JProGraM.html.

6. New Results

6.1. Probabilistic models for large graph

We have developped new approaches for the statistical analysis of large-scale undirected graphs. The main insight is to exploit the spectral decomposition of subgraph samples, and in particular their Fiedler eigenvalues, as basic features for density estimation and probabilistic inference. Our contributions are twofold. First, we develop a conditional random graph model for learning to predict links in information networks (such as scientific coauthorship and email communication). Second, we propose to apply the resulting model to graph generation and link prediction. This work is to published in the *Journal of Machine Learning Research*, the top journal in the field of machine learning.

6.2. Learning in hypergraphs

In this work, we focus on the problem of learning from several sources of heterogeneous data represented as input graphs that encode different relations over the same set of nodes. Our goal is to merge those input graphs by embedding them into an Euclidean space related to the commute time distance in the original graphs. Our algorithm output a combined kernel that can be used for different graph learning tasks. This work has been published in [5].

The approach designed in that paper has raised a new definition of undirected hypergraphs with bipartite hyperedges. A bipartite hyperedge is a pair of disjoint sets of nodes in which every node is associated with a weight. A bipartite hyperedge can be viewed as a relation between two teams of nodes in which every node has a weighted contribution to its team. Undirected hypergraphs generalize over undirected graphs. Consistently with the case of graphs, we have studied the hypergraph spectral framework. We have defined the notions of hypergraph gradient, hypergraph Laplacian, and hypergraph kernel as the Moore-Penrose pseudoinverse of a hypergraph Laplacian. Therefore, smooth labeling of (teams of) nodes and hypergraph regularization methods can be performed. Contrary to the graph case, we show that the class of hypergraph Laplacians is closed by the pseudoinverse operation (thus it is also the class of hypergraphs kernels), and is closed by convex linear combination. closure properties allow us to define (hyper)graph combinations and operations while keeping a hypergraph interpretation of the result. We exhibit a subclass of signed graphs that can be associated with hypergraphs in a constructive way. A hypergraph and its associated signed graph have the same Laplacian. This property allows us to define a distance between nodes in undirected hypergraphs as well as in the subclass of signed graphs. The distance coincides with the usual definition of commute-time distance when the equivalent signed graph turns out to be a graph. We claim that undirected hypergraphs open the way to solve new learning tasks and model new problems based on set similarity or dominance. We are currently exploring applications for modeling games between teams and for graph summarization. This work [8] has been submitted to *Journal of Machine Learning Research*.

6.3. Natural Language Processing

In [7] and [3], we develop a new algorithm for drastically improving a pairwise coreference classification system. Specifically, this algorithm works by learning the best partition over mention type pairs by training different pairwise coreference models for each pair type. In effect, our algorithm finds the optimal feature space (from a base feature set and set of types) for separating coreferential mention pairs, but it remains tractable by exploiting the structure of the hierarchies built from the pair types. In [6], we propose a new approach for the automatic identification of so-called implicit discourse relations. Our system combines hand labeled examples and automatically annotated examples (based on explicit relations) using different methods inspired by work on domain adapation. Our system is evaluated empirically and yields important performance gains compared to only using hand-labeled data. This paper has received the best paper award at the *TALN 2013* conference, the national NLP conference.

6.4. Query Induction

We have proposed a new algorithm for query learning that combines schema-guided pruning heuristics with the traditional learning algorithm for tree automata from positive and negative examples. We show that this algorithm is justified by a formal learning model, and that for stable queries it performs very well in practice of XML information extraction. This work [1] has alo been published in *JMLR*.

6.5. Learning Transducers

We have pursued the work on learning finite state tree-to-word transducers. Tree-to-word transformations are ubiquitous in computer science. They are the core of many computation paradigms from the evaluation of abstract syntactic trees to modern programming languages *XSLT*. We have extended the results obtained last year on the study of a class of sequential top-down tree-to-word transducers, called *STWs*. Transducers in *STWs* are capable of: concatenation in the output, producing arbitrary context-free languages, deleting inner nodes, and verifying that the input tree belongs to the domain even when deleting parts of it. These features are often missing in tree-to-tree transducers, and for instance, make *STWs* incomparable with the class of top-down tree-to-tree transducers. The class of *STWs* has several interesting properties, in particular we proposed in 2011 a normal for *STWs*.

In [4], we present a Myhill-Nerode characterization of the corresponding class of sequential tree-to-word transformations. Next, we investigate what learning of *STWs* means, identify fundamental obstacles, and propose a learning model with abstain. Finally, we present a polynomial learning algorithm.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

First, we are involved in the HERMES project along a collaboration with the SEQUEL INRIA team and with a consortium of companies. In that collaboration, the envisioned applications is the design of recommender systems for commercial data. One objective is to provide social recommendations, that is to take into accounts in the recommendations, social relationships between users and the content of messages posted by users in forums.

Second, we start a one to one cooperation with the CLIC AND WALK company along the PhD thesis of PAULINE WAUQUIER. The company 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,... We will study the problem of adaptive graph construction from the clicwalkers network. Node (users) classification and clustering algorithms will be applied. For the problem of survey recommendations, the problem of teams constitution in a bipartite graphs of users and surveys will be studied. Random graph modeling and generative models of random graphs will be one step towards the prediction of the evolution of clicwalkers community.

Third, we have started a transfer collaboration with the MUSIC STORY company. In a first phase, we have considered the question of collecting musical metadata from heterogeneous sources. We have proposed machine learning methods and similarity measures for curating metadata. The MUSIC STORY company has close industrial collaborations with the DEEZER company. Current discussions between MAGNET and these two companies are open on social recommender systems for music.

Last, we work with physicians at the Lille hospital (CHRU) on the detection of brain anomalies related to epilepsy. Hence, we will use connectome data which is an approximate map of neural connections at different scales. The connectome can be modeled by a weighted graph. Available data include graphs constructed at different times for a given patient, also graphs for healthy patients and epileptic patients. One objective of the research project is to study how the connectome together with other signals, like functional magnetic resonance imaging (FRMI), MEG and EEG can be efficiently combined in order to detect abnormal brain regions. We will consider diffusion algorithms in graphs to test whether diffusion processes in the brain can be explained with the connectome. We will also consider learning algorithms related to information diffusion in order to enhance graph construction.

7.2. Bilateral Grants with Industry

7.2.1. Cifre SAP (2011-2014)

Participants: Thomas Ricatte, Gemma Casas Garriga, Marc Tommasi, Rémi Gilleron [correspondent].

GEMMA GARRIGA, and MARC TOMMASI supervise the PhD thesis (Cifre) of Thomas Ricatte together with Yannick Cras from SAP.

7.2.2. Cifre Clic and Walk (2013-2016)

Participants: Pauline Wauquier, Marc Tommasi, Mikaela Keller [correspondent].

MIKAELA KELLER and MARC TOMMASI supervise the PhD thesis (Cifre) of PAULINE WAUQUIER on graph-based recommandation together with Guillaume André from Clic and Walk.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Thèse Inria-Région NPdC (2012-2015)

Participants: Marc Tommasi [correspondent], Pascal Denis, David Chatel.

PASCAL DENIS and MARC TOMMASI supervise the PhD thesis of DAVID CHATEL on semi-supervised clustering. The PhD is funded by Inria and the "Région Nord - Pas de Calais".

8.2. National Initiatives

8.2.1. ANR

8.2.1.1. ANR Lampada (2009-2014)

Participants: Marc Tommasi [correspondent], Rémi Gilleron, Fabien Torre, Gemma Casas Garriga.

The Lampada project on “Learning Algorithms, Models and sPArse representations for structured DAta” is coordinated by Tommasi from Mostrare. Our partners are the SEQUEL project of Inria Lille Nord Europe, the LIF (Marseille), the HUBERT CURIEN laboratory (Saint-Etienne), and LIP6 (Paris). More information on the project can be found on http://lampada.gforge.inria.fr/.

8.2.2. Competitivity Clusters

We are part of FUI HERMES (2012-2015), a joint project in collaboration with many companies (Auchan, KeyneSoft, Cylande, ...). The main objective is to develop a platform for contextual customer relation management. The project started in November 2012.

8.3. European Initiatives

8.3.1. Collaborations in European Programs, except FP7

Program: ERC Advanced Grant

Project acronym: STAC

Project title: Strategic conversation

Duration: Sept. 2011 - Aug. 2016

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

Other partners: School of Informatics, Edinburgh University; Heriot Watt University, Edinburgh

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.

In addition, MAGNET, in collaboration with SEQUEL, is part of the INRIA Lille - Nord Europe site for the European Network of Excellence in Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL2).

9. Dissemination

9.1. Scientific Animation

9.1.1. Invited Talks

FABIO VITALE was invited to give a talk at the Computer Science department of Copenhagen University in November 2013 on the topic of *Machine Learning on Trees and Graphs*.

9.1.2. Program Committees

PASCAL DENIS served as member of the program committee of ACL 2013, EACL 2013, EMNLP 2013, *SEM 2013.

ANTONINO FRENO served as member of the program committee of ECML-PKDD 2013.

MIKAELA KELLER was reviewer for ECML-PKDD 2013.

FABIEN TORRE served as member of the program committee of IClaNov workshop at ICDM 2013, ICPRAM 2014, EGC 2014, CluCo workshop at EGC 2014.

MARC TOMMASI served as member of the program committee of CAP 2013 and CIAA 2013.

9.1.3. Hiring Committees

MIKAELA KELLER served as member of the hiring committee for a MdC position at University of Calais. and as member of the hiring committee for a MdC position at University of Lille 3.

RÉMI GILLERON served as member of the hiring committee for a MdC position at University of Saint Etienne.

MARC TOMMASI served as member of the hiring committee for a MdC position at University of Lille 3 and as member of the hiring committee for a Professor position at University of Paris 6.

9.1.4. Other Committees

PASCAL DENIS served as reviewer for the French National Research Agency (ANR), Programme Blanc.

MIKAELA KELLER served as reviewer for the French National Research Agency (ANR), Programme Blanc.

FABIEN TORRE is elected member of Conseil national des universités (Section 27).

MARC TOMMASI served as member of the SIMI2 committee of the French National Research Agency (ANR). He is also coordinator for the ANR Lampada project.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Master MOCAD: PASCAL DENIS, Extraction d'information, 18h, M2, Université Lille 1, France

Master MOCAD: ANTONINO FRENO, Extraction d'information, 16h, M2, Université Lille 1, France

RÉMI GILLERON is in charge of master MIASHS at Université Lille 3, France

Master MIASHS: RÉMI GILLERON, Classification supervisée, 36h, M1, Université Lille 3, France

Master MIASHS: RÉMI GILLERON, Classification non supervisée, 24h, M1, Université Lille 3, France

Master MIASHS: RÉMI GILLERON, Recherche d'information, 24h, M1, Université Lille 3, France

Master MIASHS: RÉMI GILLERON, Web sémantique, 36h, M2, Université Lille 3, France

Master MIASHS: RÉMI GILLERON, Programmation R, 24h, M1, Université Lille 3, France

Licence EMO: RÉMI GILLERON, Cours bases de données, 18h, L3, Université Lille 3, France

Licence MIASHS: RÉMI GILLERON, Programmation Python, 36h, L1, Université Lille 3, France

Licence: MIKAELA KELLER, Apprentissage statistique, 63h, L3, Université Lille 3, France

Master LTTAC: FABIEN TORRE, Traitements automatiques des textes, 55.5h, M1, Université Lille 3, France

Master IDEMM: FABIEN TORRE, Langages du web, 37.5h, M2, Université Lille 3, France

Master GIDE: FABIEN TORRE, Algorithmique et programmation PHP pour le web, 75h, M2, Université Lille 3, France

Master ICDD: FABIEN TORRE, Internet et ses langages, 37h, M1, Université Lille 3, France

Master ID: MARC TOMMASI, Réseaux, 60h, M1, Université Lille 3, France

Master ICCD: MARC TOMMASI, Représentation et codage de l'information, 37h, M1, Université Lille 3, France

9.2.2. Supervision

PhD in progress: THOMAS RICATTE, Learning in Multiple Graphs with Hypergraphs, Université Lille 3, since Sept. 2011, Marc TommasiLille, scheduled for June 2014, MARC TOMMASI, RÉMI GILLERON, and GEMMA GARRIGA

PhD in progress: DAVID CHATEL, Supervised Spectral Clustering and Information Diffusion in Graphs of Texts, Université Lille 1, since Sept. 2012, PASCAL DENIS and MARC TOMMASI

PhD in progress: PAULINE WAUQUIER, Recommendation in Information Networks, Université Lille 1, since Dec. 2013 supervised by MARC TOMMASI and MIKAELA KELLER

PhD in progress: AHMAD HOSSEINI, Machine Learning for Information Diffusion in Graphs, Université Lille 1, since Dec. 2013, MARC TOMMASI and MIKAELA KELLER and PHILIPPE PREUX

PhD in progress: Grégoire Laurence, Learning Tree Transducers, Université Lille 1, since Sept. 2008, MARC TOMMASI and JOACHIM NIEHREN

PhD in progess: Jean Decoster, Relation learning for XML documents, Université Lille 1, since Sept. 2009, FABIEN TORRE and RÉMI GILLERON

PhD in progress: EMMANUEL LASSALLE, Improved Coreference Resolution with Feature Space Learning, Université Paris-Diderot, since Sept. 2010, PASCAL DENIS and LAURENCE DANLOS (Université Paris-Diderot)

PhD in progress: CHLOÉ BRAUD, Discourse Relation Identification from Labeled and Unlabeled Data, Université Paris-Diderot, since Sept. 2011, PASCAL DENIS and LAURENCE DANLOS (Université Paris-Diderot)

9.2.3. Juries

RÉMI GILLERON was member of the PhD committee of Emilie Morvant, Université de Marseille, France

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] J. NIEHREN, J. CHAMPAVÈRE, R. GILLERON, A. LEMAY. *Query Induction with Schema-Guided Pruning Strategies*, in "Journal of Machine Learning Research", April 2013, vol. 14, pp. 927–964, http://hal.inria.fr/inria-00607121

International Conferences with Proceedings

[2] V. ANTOINE, N. ASHER, P. MULLER, P. DENIS, S. AFANTENOS. *Expressivity and comparison of models of discourse structure*, in "Special Interest Group on Discourse and Dialogue", Metz, France, August 2013, http://hal.inria.fr/hal-00838260

[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, 2013, http://hal.inria.fr/hal-00838192

[4] G. LAURENCE, A. LEMAY, J. NIEHREN, S. STAWORKO, M. TOMMASI. *Learning Sequential Tree-to-Word Transducers*, in "8th International Conference on Language and Automata Theory and Applications", Madrid, Spain, Springer, March 2014, http://hal.inria.fr/hal-00912969

[5] T. RICATTE, G. GARRIGA, R. GILLERON, M. TOMMASI. *Learning from Multiple Graphs using a Sigmoid Kernel*, in "The 12th International Conference on Machine Learning and Applications (ICMLA'13)", Miami, United States, December 2013, http://hal.inria.fr/hal-00913237

National Conferences with Proceedings

[6] C. BRAUD, P. DENIS. *Identification automatique des relations discursives "implicites" à partir de données annotées et de corpus bruts*, in "TALN - 20ème conférence du Traitement Automatique du Langage Naturel 2013", Sables d'Olonne, France, June 2013, vol. 1, pp. 104-117, http://hal.inria.fr/hal-00830983

[7] E. LASSALLE, P. DENIS. *Apprentissage d'une hiérarchie de modèles à paires spécialisés pour la résolution de la coréférence*, in "TALN 2013 - 20ème conférence du Traitement Automatique du Langage Naturel 2013", Les Sables-d'Olonne, France, June 2013, http://hal.inria.fr/hal-00825617

Research Reports

[8] T. RICATTE, G. GARRIGA, R. GILLERON, M. TOMMASI. , *A Spectral Framework for a Class of Undirected Hypergraphs*, December 2013, http://hal.inria.fr/hal-00914286

Other Publications

[9] A. FRENO, M. KELLER, M. TOMMASI. , *Probability Estimation over Large-Scale Random Networks via the Fiedler Delta Statistic*, December 2013, http://hal.inria.fr/hal-00922432

References in notes

[10] 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

[11] 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

[12] M. BELKIN, P. NIYOGI. *Towards a Theoretical Foundation for Laplacian-Based Manifold Methods*, 2008, vol. 74, n^{o} 8, pp. 1289-1308

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

[14] 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

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

[16] N. CESA-BIANCHI, C. GENTILE, F. VITALE, G. ZAPPELLA. *A Linear Time Active Learning Algorithm for Link Classification*, in "Proc of NIPS", 2012, pp. 1619-1627

[17] 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

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

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

[20] A. FRENO, G. C. GARRIGA, M. KELLER. *Learning to Recommend Links Using Graph Structure and Node Content*, in "NIPS Workshop on Choice Models and Preference Learning", 2011

[21] A. FRENO, M. KELLER, C. GARRIGA, M. TOMMASI. *Spectral Estimation of Conditional Random Graph Models for Large-Scale Network data*, in "Proc. of UAI 2012", Avalon, États-Unis, 2012, http://hal.inria.fr/hal-00714446

[22] 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

[23] 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

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

[25] M. HERBSTER, S. PASTERIS, F. VITALE. *Online Sum-Product Computation Over Trees*, in "Proc. of NIPS", 2012, pp. 2879-2887

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

[27] 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

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

[29] 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

[30] 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

[31] 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

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

[33] 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

[34] 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

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