RESEARCH CENTRE

Saclay - Île-de-France

IN PARTNERSHIP WITH:

Ecole Polytechnique

2020 ACTIVITY REPORT

Project-Team CEDAR

Rich Data Exploration at Cloud Scale

IN COLLABORATION WITH: Laboratoire d'informatique de l'école polytechnique (LIX)

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Contents

Project-Team CEDAR

Creation of the Team: 2016 January 01, updated into Project-Team: 2018 April 01

Keywords

Computer sciences and digital sciences

A3.1.1. – Modeling, representation

A3.1.2. – Data management, quering and storage

A3.1.3. – Distributed data

A3.1.6. – Query optimization

A3.1.7. – Open data

A3.1.8. – Big data (production, storage, transfer)

A3.1.9. – Database

A3.2.1. – Knowledge bases

A3.2.3. – Inference

A3.2.4. – Semantic Web

A3.2.5. – Ontologies

A3.3.1. – On-line analytical processing

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A3.4.1. – Supervised learning

A3.4.6. – Neural networks

A3.4.8. – Deep learning

A9.1. – Knowledge

A9.2. – Machine learning

Other research topics and application domains

B8.5.1. – Participative democracy

B9.5.6. – Data science

B9.7.2. – Open data

1 Team members, visitors, external collaborators

Research Scientists

- Ioana Manolescu [Team leader, Inria, Senior Researcher, HDR]
- Angelos Anadiotis [École polytechnique, Researcher]
- Oana Balalau [Inria, Starting Research Position]
- Yanlei Diao [École polytechnique, Researcher]

Post-Doctoral Fellows

- Mirjana Mazuran [Inria, until Oct 2020]
- Fei Song [École polytechnique]

PhD Students

- Nelly Barret [Inria, from Oct 2020]
- Maxime Buron [Inria, until Oct 2020]
- Luciano Di Palma [École polytechnique]
- Qi Fan [École polytechnique]
- Pawel Guzewicz [École polytechnique]
- Enhui Huang [École polytechnique]
- Vincent Jacob [École polytechnique]
- Khaled Zaouk [École polytechnique]
- Fanzhi Zhu [Institut Polytechnique de Paris, from Jul 2020]

Technical Staff

- Mhd Yamen Haddad [Inria, Engineer, from Mar 2020 until Jun 2020]
- Jean Langlois-Berthelot [Inria, Engineer, from May 2020 until Jul 2020]
- Moustafa Latrache [Inria, Engineer]
- Tayeb Merabti [Inria, Engineer]
- Saumya Sahai [Inria, Engineer, from Jul 2020 until Aug 2020]
- Arnab Sinha [École polytechnique, Engineer]

Interns and Apprentices

- Theo Bouganim [Inria, Apprentice, from Oct 2020]
- Irene Burger [Inria, from May 2020 until Aug 2020]
- Francesco Chimienti [Inria, from Sep 2020]
- Fanch Morvan [Inria, from May 2020 until Aug 2020]
- Guillaume Thiry [Inria, from May 2020 until Aug 2020]
- Jingmao You [Inria, from May 2020 until Sep 2020]
- Youssr Youssef [Inria, from May 2020 until Aug 2020]
- Xin Zhang [Inria, from Feb 2020 until Aug 2020]

Administrative Assistant

- Alexandra Merlin [Inria, from Oct 2020]

External Collaborators

- Rana Alotaibi [UCSD]
- Irene Burger [École polytechnique, until May 2020]
- Gauthier Guinet [École polytechnique, until Aug 2020]
- Mhd Yamen Haddad [Oracle - Zurich, from Jul 2020]
- Stephane Horel [Le Monde, from Nov 2020]
- Roxana Horincar [Thales, from May 2020]
- Julien Leblay [National Institute of Advanced Industrial Science and Technology-Japan, until Jun 2020]
- Catarina Pinto Conceicao [Institut Supérieur Technique - Lisbonne, from Feb 2020]
- Saumya Sahai [Université d'État de l'Ohio - Columbus USA, NC, from May 2020 until Jun 2020]
- Jingmao You [École polytechnique, until May 2020]
- Youssr Youssef [École nationale de la statistique et de l'administration économique Paris, from Feb 2020]

2 Overall objectives

Our research aims at **models, algorithms and tools for highly efficient, easy-to-use data and knowledge management**; throughout our research, **performance at scale** is a core concern, which we address, among other techniques, by designing algorithms for a **cloud (massively parallel)** setting. Our scientific contributions fall in three interconnected areas:

Expressive models for new applications As data and knowledge applications keep extending to novel application areas, we work to devise appropriate data and knowledge models, endowed with formal semantics, to capture such applications' needs. This work mostly concerns the domains of data journalism and journalistic fact checking;

Optimization and performance at scale This topic is at the heart of Y. Diao's ERC project "Big and Fast Data", which aims at optimization with performance guarantees for real-time data processing in the cloud. Machine learning techniques and multi-objectives optimization are leveraged to build performance models for data analytics the cloud. The same boal is shared by our work on efficient evaluation of queries in dynamic knowledge bases.

Data discovery and exploration Today's Big Data is complex; understanding and exploiting it is difficult. To help users, we explore: compact summaries of knowledge bases to abstrac their structure and help users formulate queries; interactive exploration of large relational databases; techniques for automatically discovering interesting information in knowledge bases; and keyword search techniques over Big Data sources.

3 Research program

3.1 Scalable Heterogeneous Stores

Big Data applications increasingly involve *diverse* data sources, such as: structured or unstructured documents, data graphs, relational databases etc. and it is often impractical to load (consolidate) diverse data sources in a single repository. Instead, interesting data sources need to be exploited "as they are", with the added value of the data being realized especially through the ability to combine (join) together data from several sources. Systems capable of exploiting diverse Big Data in this fashion are usually termed *polystores*. A current limitation of polystores is that data stays captive of its original storage system, which may limit the data exploitation performance. We work to devise highly efficient storage systems for heterogeneous data across a variety of data stores.

3.2 Semantic Query Answering

In the presence of data semantics, query evaluation techniques are insufficient as they only take into account the database, but do not provide the reasoning capabilities required in order to reflect the semantic knowledge. In contrast, (ontology-based) query answering takes into account both the data and the semantic knowledge in order to compute the full query answers, blending query evaluation and semantic reasoning.

We aim at designing efficient semantic query answering algorithms, both building on cost-based reformulation algorithms developed in the team and exploring new approaches mixing materialization and reformulation.

3.3 Multi-Model Querying

As the world's affairs get increasingly more digital, a large and varied set of data sources becomes available: they are either structured databases, such as government-gathered data (demographics, economics, taxes, elections, ...), legal records, stock quotes for specific companies, un-structured or semi-structured, including in particular graph data, sometimes endowed with semantics (see e.g. the Linked Open Data cloud). Modern data management applications, such as data journalism, are eager to combine in innovative ways both static and dynamic information coming from structured, semi-structured, and un-structured databases and social feeds. However, current content management tools for this task are not suited for the task, in particular when they require a lenghy rigid cycle of data integration and consolidation in a warehouse. Thus, we see a need for flexible tools allowing to interconnect various kinds of data sources and to query them together.

3.4 Interactive Data Exploration at Scale

In the Big Data era we are faced with an increasing gap between the fast growth of data and the limited human ability to comprehend data. Consequently, there has been a growing demand of data management tools that can bridge this gap and help users retrieve high-value content from data more effectively. To

respond to such user information needs, we aim to build interactive data exploration as a new database service, using an approach called “explore-by-example”.

3.5 Exploratory Querying of Semantic Graphs

Semantic graphs including data and knowledge are hard to apprehend for users, due to the complexity of their structure and oftentimes to their large volumes. To help tame this complexity, in prior research (2014), we have presented a full framework for RDF data warehousing, specifically designed for heterogeneous and semantic-rich graphs. However, this framework still leaves to the users the burden of chosing the most interesting warehousing queries to ask. More user-friendly data management tools are needed, which help the user discover the interesting structure and information hidden within RDF graphs. This research has benefitted from the arrival in the team of Mirjana Mazuran, as well as from the start of the PhD thesis of Paweł Guzewicz, co-advised by Yanlei Diao and Ioana Manolescu.

3.6 An Unified Framework for Optimizing Data Analytics

Data analytics in the cloud has become an integral part of enterprise businesses. Big data analytics systems, however, still lack the ability to take user performance goals and budgetary constraints for a task, collectively referred to as task objectives, and automatically configure an analytic job to achieve the objectives.

Our goal, is to come up with a data analytics optimizer that can automatically determine a cluster configuration with a suitable number of cores as well as other runtime system parameters that best meet the task objectives. To achieve this, we also need to design a multi-objective optimizer that constructs a Pareto optimal set of job configurations for task-specific objectives, and recommends new job configurations to best meet these objectives.

3.7 Elastic resource management for virtualized database engines

Database engines are migrating to the cloud in order to leverage the opportunities for efficient resource management by adapting to the variations and the heterogeneity of the workloads. Resource management in a virtualized setting, like the cloud, need to be enforced in a performance-efficient manner in order to avoid introducing overheads to the execution.

We design elastic systems which change their configuration at runtime with minimal cost in order to adapt to the workload every time. Changes in the design include both different resource allocation and different data layouts. We consider different workloads including transactional, analytical and mixed and we study the performance implications on different configurations in order to eventually propose a set of adaptive algorithms.

4 Application domains

4.1 Cloud Computing

Cloud computing services are strongly developing and more and more companies and institutions resort to running their computations in the cloud, in order to avoid the hassle of running their own infrastructure. Today’s cloud service providers guarantee machine availabilities in their Service Level Agreement (SLA), without any guarantees on performance measures according to a specific cost budget. Running analytics on big data systems require the user not to only reserve the suitable cloud instances over which the big data system will be running, but also setting many system parameters like the degree of parallelism and granularity of scheduling. Chosing values for these parameters, and chosing cloud instances need to meet user objectives regarding latency, throughput and cost measures, which is a complex task if it’s done manually by the user. Hence, we need need to transform cloud service models from availabily to user performance objective rises and leads to the problem of multi-objective optimization. Research carried out in the team within the ERC project “Big and Fast Data Analytics” aims to develop a novel optimization framework for providing guarantees on the performance while controlling the cost of data processing in the cloud.

4.2 Computational Journalism

Modern journalism increasingly relies on content management technologies in order to represent, store, and query source data and media objects themselves. Writing news articles increasingly requires consulting several sources, interpreting their findings in context, and crossing links between related sources of information. CEDAR research results directly applicable to this area provide techniques and tools for rich Web content warehouse management. Within the ANR ContentCheck project, and following through the SourcesSay AI Chair, we work to devise concrete algorithms and platforms to help journalists perform their work better and/or faster. This work is in collaboration with journalists form Le Monde's: its fact-checking team "Les Décodeurs" was involved in ANR ContentCheck, while our current work continues in collaboration with award-winning investigative journalist Stéphane Horel (https://fr.wikipedia.org/wiki/St%C3%A9phane_Horel)

5 Highlights of the year

5.1 Awards

Ioana Manolescu has been a co-recipient of the **ACM SIGMOD 2020 Contributions Award** for "innovative work in the data management community to encourage scientific reproducibility of our publications". According to the award notice (https://sigmod.org/2020-sigmod-contributions-award/): "Reproducibility was introduced at the 2008 SIGMOD Conference" (when Ioana Manolescu has been the first SIGMOD Repeatability chair) "and since then has influenced how the community approaches experimental evaluation. It has also influenced similar efforts within ACM".

Maxime Buron has been awarded the **Prix de thèse de la communauté BDA 2020**, rewarding his PhD thesis work.

6 New software and platforms

6.1 New software

6.1.1 AIDES

Keywords: Data Exploration, Active Learning

Functional Description: AIDES is a data exploration software. It allows a user to explore a huge (tabular) dataset and discover tuples matching his or her interest. Our system repeatedly proposes the most informative tuples to the user, who must annotate them as "interesting" / "not-interesting", and as iterations progress an increasingly accurate model of the user's interest region is built. Our system also focuses on supporting low selectivity, high-dimensional interest regions.

Contacts: Yanlei Diao, Enhui Huang, Luciano Di Palma

6.1.2 OntoSQL

Keywords: RDF, Semantic Web, Querying, Databases

Functional Description: OntoSQL is a tool providing three main functionalities: - Loading RDF graphs (consisting of data triples and possibly a schema or ontology) into a relational database, - Saturating the data based on the ontology. Currently, RDF Schema ontologies are supported. - Querying the loaded data using conjunctive queries. Data can be loaded either from distinct files or from a single file containing them both. The loading process allows to choose between two storage schemas: - One triples table. - One table per role and concept. Querying provides an SQL translation for each conjunctive query according to the storage schema used in the loading process, then the SQL query is evaluated by the underlying relational database.

URL: https://ontosql.inria.fr/

Contacts: Ioana Manolescu, Tayeb Merabti

Participants: Ioana Manolescu, Michaël Thomazo, Tayeb Merabti

Partner: Université de Rennes 1

6.1.3 ConnectionLens

Keywords: Data management, Big data, Information extraction, Semantic Web

Functional Description: ConnectionLens treats a set of heterogeneous, independently authored data sources as a single virtual graph, whereas nodes represent fine-granularity data items (relational tuples, attributes, key-value pairs, RDF, JSON or XML nodes...) and edges correspond either to structural connections (e.g., a tuple is in a database, an attribute is in a tuple, a JSON node has a parent...) or to similarity (sameAs) links. To further enrich the content journalists work with, we also apply entity extraction which enables to detect the people, organizations etc. mentioned in text, whether full-text or text snippets found e.g. in RDF or XML. ConnectionLens is thus capable of finding and exploiting connections present across heterogeneous data sources without requiring the user to specify any join predicate.

URL: https://team.inria.fr/cedar/connectionlens/

Publications: hal-02934277, hal-02904797, hal-01841009

Authors: Tayeb Merabti, Helena Galhardas, Julien Leblay, Ioana Manolescu, Oana Balalau, Catarina Pinto Conceicao

Contact: Manolescu Ioana

6.1.4 INSEE-Search

Keywords: Document ranking, RDF

Functional Description: Searching for relevant data cells (or data row/column) given a query in natural language (French)

Publications: hal-01915148, hal-01745768

Contacts: Ioana Manolescu, Xavier Tannier, Tien Duc Cao

Participants: Ioana Manolescu, Xavier Tannier, Tien Duc Cao

6.1.5 RDFQuotient

Name: Quotient summaries of RDF graphs

Keywords: RDF, Graph algorithmics, Graph visualization, Graph summaries, Semantic Web

Functional Description: RDF graphs can be large and heterogeneous, making it hard for users to get acquainted with a new graph and understand whether it may have interesting information. To help users figure it out, we have devised novel equivalence relations among RDF nodes, capable of recognizing them as equivalent (and thus, summarize them together) despite the heterogeneity often exhibited by their incoming and outgoing node properties. From these relations, we derive four novel summaries, called Weak, Strong, Typed Weak and Typed Strong, and show how to obtain from them compact and enticing visualizations.

Publications: hal-01325900v6, hal-01808737

Authors: Ioana Manolescu, Pawel Guzewicz, François Goasdoué

Contact: Manolescu Ioana

Participants: Ioana Manolescu, Pawel Guzewicz, François Goasdoué

Partner: Université de Rennes 1

6.1.6 AIDEme

Keywords: Active Learning, Data Exploration

Scientific Description: AIDEme is a large-scale interactive data exploration system that is cast in a principled active learning (AL) framework: in this context, we consider the data content as a large set of records in a data source, and the user is interested in some of them but not all. In the data exploration process, the system allows the user to label a record as "interesting" or "not interesting" in each iteration, so that it can construct an increasingly-more-accurate model of the user interest. Active learning techniques are employed to select a new record from the unlabeled data source in each iteration for the user to label next in order to improve the model accuracy. Upon convergence, the model is run through the entire data source to retrieve all relevant records.

A challenge in building such a system is that existing active learning techniques experience slow convergence in learning the user interest when such exploration is performed on large datasets: for example, hundreds of labeled examples are needed to learn a user interest model over 6 attributes, as we showed using a digital sky survey of 1.9 million records. AIDEme employs a set of novel techniques to overcome the slow convergence problem:

• Factorization: We observe that a user labels a data record, her decision making process often can be broken into a set of smaller questions, and the answers to these questions can be combined to derive the final answer. This insight, formally modeled as a factorization structure, allows us to design new active learning algorithms, e.g., factorized version space algorithms [2], that break the learning problem into subproblems in a set of subspaces and perform active learning in each subspace, thereby significantly expediting convergence.

• Optimization based on class distribution: Another interesting observation is that when projecting the data space for exploration onto a subset of dimensions, the user interest pattern projected onto such a subspace often entails a convex object. When such a subspatial convex property holds, we introduce a new "dual-space model" (DSM) that builds not only a classification model from labeled examples, but also a polytope model of the data space that offers a more direct description of the areas known to be positive, areas known to be negative, and areas with unknown labels. We use both the classification model and the polytope model to predict unlabeled examples and choose the best example to label next. • Formal results on convergence: We further provide theoretical results on the convergence of our proposed techniques. Some of them can be used to detect convergence and terminate the exploration process. • Scaling to large datasets: In many applications the dataset may be too large to fit in memory. In this case, we introduce subsampling procedures and provide provable results that guarantee the performance of the model learned from the sample over the entire data source.

Functional Description: There is an increasing gap between fast growth of data and limited human ability to comprehend data. Consequently, there has been a growing demand for analytics tools that can bridge this gap and help the user retrieve high-value content from data. We introduce AIDEme, a scalable interactive data exploration system for efficiently learning a user interest pattern over a large dataset. The system is cast in a principled active learning (AL) framework, which iteratively presents strategically selected records for user labeling, thereby building an increasingly-more-accurate model of the user interest. However, a challenge in building such a system is that existing active learning techniques experience slow convergence when learning the user interest on large datasets. To overcome the problem, AIDEme explores properties of the user labeling process and the class distribution of observed data to design new active learning algorithms, which come with provable results on model accuracy, convergence, and approximation, and have evaluation results showing much improved convergence over existing AL methods while maintaining interactive speed.

Release Contributions: Project code can be found over: https://gitlab.inria.fr/ldipalma/aideme

URL: http://www.lix.polytechnique.fr/aideme

Contacts: Yanlei Diao, Luciano Di Palma, Enhui Huang

Participants: Luciano Di Palma, Enhui Huang

6.1.7 ConnectionLensInMem

Keywords: Data management, Graph processing

Functional Description: In-memory graph-based keyword search. It works in collaboration with ConnectionLens and it focuses on parallelization of the query execution. The software includes a module to export a graph from ConnectionLens PostgreSQL warehouse which can then be loaded in the main memory for querying.

Contact: Angelos Anadiotis

7 New results

7.1 Querying and analyzing semantic graphs

7.1.1 Efficient query answering over semantic graphs

In this area, the end of the PhD thesis of M. Buron has lead to two main contributions.

The first is related to the **performance of query answering in RDF databases**. This strongly depends on the data *layout*, that is, the way data is split in persistent data structures. In [15], we consider answering Basic Graph Pattern Queries (BGPQs), and in particular those with variables (also) in class and property positions, in the presence of RDFS ontologies, both through data saturation and query reformulation. We show that such demanding queries often lead to inefficient query answering on two popular storage layouts, so-called T and CP. We present *novel query answering algorithms on the TCP layout*, which combines T and CP. In exchange to occupying more storage space, e.g. on an inexpensive disk, TCP avoids the bad or even catastrophic performance that T and/or CP sometimes exhibit.

The second and main contribution concerns the **efficient integration of heterogeneous, structured databases, as semantic graphs**. The proliferation of heterogeneous data sources in many application contexts brings an urgent need for expressive and efficient data integration mechanisms. There are strong advantages to using RDF graphs as the integration format: being schemaless, they allow for flexible integration of data from heterogeneous sources; RDF graphs can be interpreted with the help of an ontology, describing application semantics; last but not least, RDF enables joint querying of the data and the ontology.

To address this need, in [17], we formalize *RDF Integration Systems (RIS)*, Ontology Based-Data Access mediators, that go beyond the state of the art in the ability to expose, integrate and flexibly query data from heterogeneous sources through GLAV (global-local-as-view) mappings. We devise several *query answering strategies*, based on an innovative integration of LAV view-based rewriting and a form of mapping saturation. Our experiments show that one of these strategies brings strong performance advantages, resulting from a balanced use of mapping saturation and query reformulation. A demonstration of the system implementing these techniques has been shown at the VLDB 2020 conference [16].

7.1.2 Quotient summaries of RDF graphs

To help users get familiar with large RDF graphs, we can use RDF summarization techniques. We study quotient summaries of RDF graphs, that is, graph summaries derived from a notion of equivalence among RDF graph nodes. In our recently published work [11], we make the following contributions: (i) four novel summaries, which are often small and easy-to-comprehend, in the style of E-R diagrams; (ii) efficient (amortized linear-time) algorithms for computing these summaries either from scratch or incrementally, reflecting additions to the graph; (iii) the first formal study of the interplay between RDF graph saturation in the presence of an RDFS ontology, and summarization; we provide a sufficient condition for a highly efficient shortcut method to build the quotient summary of a graph without saturating it; (iv) formal results establishing the shortcut conditions for some of our summaries and others from the literature; (v) experimental validations of our claim within a tool available online.

This research has also been featured in an invited keynote [19].

7.2 Data management for analysing digital arenas

7.2.1 Graph integration of heterogeneous data sources for data journalism

Work carried within the ANR ContentCheck and especially the ANR AI Chair SourcesSay projects has focused on developing a platform for integrating arbitrary heterogeneous data into a graph, then exploring and querying that graph in a simple intuitive manner through keyword search. The main technical challenges are: (i) how to interconnect structured and semistructured data sources? We address this th ough information extraction (when an entity appears in two data sources, or two places in the same graph, we only create one node, thus interlinking the two locations), and through similarity comparisons; (ii) how to find all connections between nodes matching certain search criteria, or certain keywords? The question is particularly challenging in our context since ConnectionLens graphs can be quite large and query answers can traverse edges in both directions.

Intense work has been invested in the ConnectionLens prototype, in particular thanks to the recruitment of T. Merabti as an AI engineer within the team [27].

- In the first part of the year, work done by Irène Burger (M1 Polytechnique student) has set the first steps toward simplifying ConnectionLens graphs through a mechanism called *abstraction* [14]. That work has been continued during the fall through the pre-doctoral contract of Nelly Barret.
- During the summer, the M2 interns Jingmao You and Youssr Youssef contributed to the development of two new modules for ConnectionLens: one for *disambiguating* named entities recognized in the text, by attaching them to known entities from the YAGO knowledge base, and the other for *converting PDF documents* into semistructured JSON (and possibly RDF) data, which allows their ingestion into ConnectionLens. Together with improvements brought to our Named Entity Recognition module and a general description of ConnectionLens graph construction, these modules were described in [30].
- As part of the internship of Mhd Yamen Haddad, we have revisited the keyword search algorithm of ConnectionLens, called GAM. We have notably introduced a new technique, called Grow2Representative, which enables to find answers for keyword queries that span over multiple datasets, yet is much more efficient than the previous technique [27] proposed for the task. The new algorithm has lead to the publication of [12].
- Finally, following the invited keynote of Ioana Manolescu at the ADBIS 2020 conference, an invited paper to a special issue of Elsevier Information Systems has been submitted [26], consolidating and unifying [30] and [12] and including larger scale experiments.

During the spring-summer of 2020, during the M1 internship of Jérémie Feitz (Polytechnique), co-supervised by E. Pietriga from the ILDA team of Inria Saclay, a new interactive user interface has been developed for querying and exploring ConnectionLens graphs. Figure 1 illustrates it on an example. This graph originates in four data sources (the four red nodes); they interconnect on common extracted entities (green and yellow nodes).

The main novelty of this interface with respect to the previous one is to support interactive exploration of the graph: users are able to explore the neighbors of a node present in the interface, interactively selecting these neighbors depending on their type etc.

This research has also lead to two invited keynotes [20, 24].

7.2.2 Novel fact-checking architectures and algorithms

Among existing sources of Open Data, statistical databases published by national and international organizations such as the International Monetary Fund, the United Nations, OECD etc. stand out for their high quality and valuable insights. However, technical means to interact easily with such sources are currently lacking.

In the past, we had worked to improve the accessibility and ease of use of statistic databases published by INSEE, the leading French statistic institute.

In 2020, during the internship of Guillaume Thiry (M1 Polytechnique), co-supervised with Leo Liberti (LIX), we have considered a new trove of statistic data, namely those structured according to

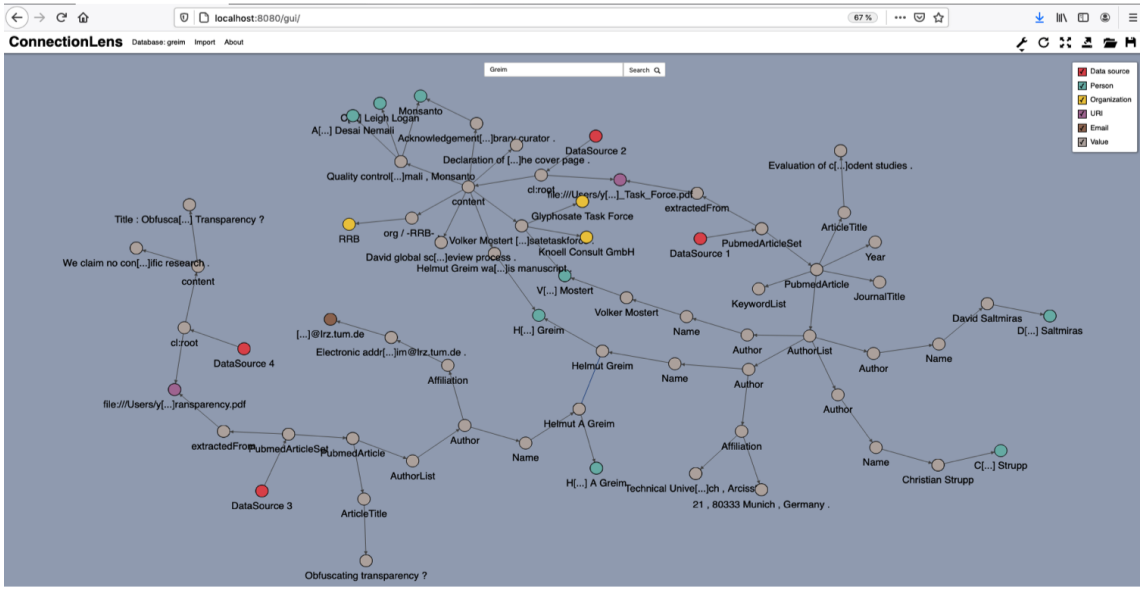

Figure 1: Sample screen shot of the new ConnectionLens interactive interface.

the SDMX standard (https://sdmx.org). SDMX is a standard to facilitate the exchange of statistical data and metadata using modern information technology. Several versions of the SDMX standard (ISO IS 17369) exist since 2004. [23] presents an effort to build an interactive Question Answering system for accessing statistical databases structured according to the SDMX standard. We describe the system architectures, its main technical choices, and present a preliminary evaluation. The system is available online (https://github.com/guillaume-thiry/OECD-Chatbot).

7.2.3 Argumentation Mining in Online Forums

In this area, we first studied propaganda dissemination in online forums and then we focused on a particular type of propaganda, fallacies. Political discussions revolve around ideological conflicts that often split the audience into two opposing parties. Both parties try to win the argument by bringing forward information. However, often this information is misleading, and its dissemination employs propaganda techniques. We analyzed the impact of propaganda on six major political forums on Reddit that target a diverse audience in two countries, the US and the UK. We focused on three research questions: who is posting propaganda? how does propaganda differ across the political spectrum? and how is propaganda received on political forums? This work was accepted for publication and will appear in 2021. Our second project was on fallacious arguments. Fallacies are generally persuasive arguments that provide insufficient or incorrect evidence that does not support the claim. We looked at the most frequent fallacies that users make on Reddit, and we presented them using the pragma-dialectical theory of argumentation. We constructed a new annotated dataset of fallacies, using user comments containing fallacy mentions as noisy labels, and cleaning the data via crowdsourcing. Finally, we studied the task of classifying fallacies using neural models. This work is under review.

7.2.4 Machine Learning for Graph Data

In this area, we worked on several topics. The first one is on how to compute subgraph embeddings in heterogeneous graphs. State-of-the-art research is focused mostly on node embeddings, with little effort dedicated to the closely related task of computing subgraph embeddings. Subgraph embeddings have many applications, such as community detection, cascade prediction, and question answering. We proposed a subgraph to subgraph proximity measure as a building block for a subgraph embedding framework in this work. Experiments on real-world datasets showed that our approach, outperformed state-of-the-art methods on several important data mining tasks. This work was published as a full paper

at PAKDD 2020 [13]. The second topic was on how to rank heterogeneous subgraphs. We focused on learning methods to rank paths and then subgraphs in heterogeneous graphs. We have initially investigated unsupervised ranking methods, such as methods based on the average personalized PageRank of nodes. We then moved to the supervised learning to rank approach. In this approach, we don't rank paths/subgraphs but meta paths/meta subgraphs. A meta path is a class of paths, where we keep the edge labels, but we replace nodes with their type. For example, Jane livesIn Paris and Jack livesIn London both belong to the metapath Person livesIn Location. In order to find the interesting meta paths, we investigated which meta paths could be used to answer popular web questions. This is ongoing work.

The last topic is on citation intent in citation networks. Many metrics have been proposed over the years to understand influential authors and influential articles. However, recent advancements in NLP have relaunched the discussion on what makes a paper influential and how a scientific field evolves over the year. In particular, recent works have looked at the intent of authors when citing other papers, highlighting six functions of a citation: citation of background work - an article relevant for the field, a motivation work which illustrates the need for the current work, an article containing a method used in the current paper, a state-of-the-art competitor, an article which the work extends, or a paper that could be an inspiration for future work. In our work we predict the intent of a citation. We depart from previous literature that took into consideration for the task only the linguistic information in an article. We incorporate more context by representing articles and citations as a heterogeneous graph, with node and edge types and labels, and predict new citations as new links in our graph. This is ongoing work.

7.3 Data exploration

7.3.1 Semantic graph exploration through interesting aggregates

As large Open Data are increasingly shared as RDF graphs today, there is a growing demand to help users discover the most interesting facets of a graph, which are often hard to grasp without automatic tools. We consider the problem of automatically identifying the k most interesting aggregate queries that can be evaluated on an RDF graph, given an integer k and a user-specified interestingness function. Our problem departs from analytics in relational data warehouses in that (i) in an RDF graph we are not given but we must identify the facts, dimensions, and measures of candidate aggregates; (ii) the classical approach to efficiently evaluating multiple aggregates breaks in the face of multi-valued dimensions in RDF data. In this work, we propose an extensible end-to-end framework that enables the identification and evaluation of interesting aggregates based on a new RDF-compatible one-pass algorithm for efficiently evaluating a lattice of aggregates and a novel early-stop technique (with probabilistic guarantees) that can prune uninteresting aggregates. Experiments using both real and synthetic graphs demonstrate the ability of our framework to find interesting aggregates in a large search space, the efficiency of our algorithms (with up to 2.9 times speedup over a similar pipeline based on existing algorithms), and scalability as the data size and complexity grow.

7.3.2 A factorized version space algorithm for interactive database exploration

In a recent trend known as "Machine Learning for Everyone", IT companies are delivering cloud platforms to help every data user to develop machine learning models for their data sets with minimum effort. A key question, however, is how to obtain a high-quality training data set for model development with minimum user effort. In other words, at the center of the new IT trend lies a critical "training data" problem. While industry solutions to this problem are limited to manual labeling or crowdsourcing, recent research on interactive data exploration (IDE) for model development bridges the gap between the data exploration and machine learning communities, and brings active learning-based data exploration to bear on the new process of model learning. In this setting, active learning is applied to select a small sequence of data instances for the user to label in order to derive an accurate model, while at the same time, offering interactive performance in presenting the next data instance for the user to review and label.

Existing active learning techniques, however, often fail to provide satisfactory performance when such models need to be built over large data sets. Not only such models often require hundreds of labeled data instances in order to reach high accuracy (slow convergence), but retrieving the next instance to label can be time consuming (inefficiency), making it incompatible with the interactive nature of the

human exploration process. To address the slow convergence and inefficiency issues, we have developed two main ideas: First, we introduce a novel version space based active learning algorithm for kernel classifiers, which not only has strong theoretical guarantees on convergence, but also allows for an efficient implementation in time and space. Second, by leveraging additional insights obtained in the user exploration and labeling process, we explore a new opportunity to factorize an active learner so that active learning can be performed in a set of low-dimensional subspaces, which further expedites convergence and reduces the user labeling effort.

More specifically, we have developed the following contributions:

1. A new theoretical framework for version space (VS) algorithms over kernel classifiers: We developed a new theoretical framework that allows for an efficient implementation of the Generalized Binary Search strategy over kernel classifiers, offering both strong theoretical guarantees on performance and an efficient implementation in time and space. We also proved generalization error bounds on accuracy and F-score, enabling our techniques to run over a sample from the original large data set with minimal performance loss.

2. Implementation and Optimizations: Based on our theoretical results, we devised an optimized VS algorithm called OptVS, which uses the hit-and-run algorithm for sampling the version space. However, hit-and-run may require thousands of iterations to output a high-quality sample, which can incur a high time cost. To reduce the cost, we develop a range of sampling optimizations to improve both sample quality and running time. In particular, we provide a highly efficient version of the rounding technique for improving the sample quality from the version space.

3. A Factorized Version Space Algorithm: Additionally, we developed a new algorithm that leverages the factorization structure provided by the user to create low-dimensional subspaces, and factorizes the version space accordingly to perform active learning in the subspaces. Compared to recent work that also used factorization for active learning, our work explores it in the new setting of VS algorithms and eliminates the strong assumptions made in prior work such as convexity of user interest patterns, resulting in significant performance improvement while increasing the applicability in real-world problems. We also managed to prove theoretical results on the optimality of our factorized VS algorithm.

Using real-world data sets and a large suite of user interest patterns, we have empirically observed that our optimized version space (VS) algorithms outperform existing VS algorithms, as well as DSM, a factorization-aware algorithm, often by a wide margin while maintaining interactive speed.

The results of our work are included in [28].

7.3.3 Learning with label noise

In active learning based data exploration, theory of active learning is applied to select a small sequence of data instances for the user to label in order to derive an accurate model, while at the same time, offering interactive performance in presenting the next data instance for the user to review and label. Several algorithms that have been proposed in the literature to address the slow convergence and inefficiency problems, making the assumption that the user-provided labels are uncorrupted. However, practical experience shows that users may mislabel some examples. Given limited labeled examples in the interactive data exploration scenario, we improve the robustness of learning algorithms in the presence of label noise by (i) applying advanced methods to collect distilled examples out of noisy examples automatically, (ii) leveraging the polytope-based model learnt on distilled examples to further filter noisy labels, and (iii) developing new sample acquisition strategies that are less sensitive to label noise. Evaluation results using real-world datasets and user interest patterns show that our proposed algorithm is far more robust than alternative algorithms and it achieves desired accuracy and efficiency under different noise levels.

7.4 Efficient Big Data Analytics

7.4.1 Scalable storage for polystores

Big data applications increasingly involve diverse datasets, conforming to different data models. Such datasets are routinely hosted in heterogeneous stores, each capable of handling one or a few data models, and each efficient for some, but not all, kinds of data processing. Systems capable of exploiting disparate data in this fashion are usually termed *polystores*. A current limitation of polystores is that applications are written taking into account which part of the data is stored in which store and how. This fails to take advantage of (i) possible redundancy, when the same data may be accessible (with different performance) from distinct data stores; (ii) previous query results (in the style of materialized views), which may be available in the stores.

The ESTOCADA system has been developed in collaboration with A. Deutsch and R. Al-Otaibi from UCSD. It proposes a novel approach that can be used in a polystore setting to transparently enable each query to benefit from the best combination of stored data and available processing capabilities. The system leverages recent advances in the area of view-based query rewriting under constraints, which we use to describe the various data models and stored data. In 2020, the system has been further enhanced with the ability to support view-based rewriting and algebraic optimizations for matrix data, at the core of modern Machine Learning data-intensive computational pipelines [9].

7.4.2 Boosting Cloud Data Analytics using Multi-Objective Optimization

In our work, we present a data analytics optimizer together with a principled multi-objective optimization approach. This approach computes a Pareto optimal set of job configurations to reveal tradeoffs between different user objectives, recommends a new job configuration that best explores such tradeoffs, and employs novel optimizations to enable such recommendations within a few seconds. Using benchmark workloads, our experimental results show that our MOO techniques outperform existing MOO methods in speed and coverage of the Pareto frontier, as well as simpler optimization methods, and can indeed recommend configurations that lead to better performance or explore interesting tradeoffs.

This work has been accepted to the proceedings of the ICDE 2021 for inclusion as a full paper [22].

7.4.3 Workload tuning using recommender systems

Spark is a widely popular massively parallel processing system which is used to execute various types of data analysis workloads. Accordingly, tuning its performance is a hard problem. In this research thread, we have casted the problem of tuning Spark workloads into a recommender systems framework.

In [29], we have introduced a representation learning architectures from three families of techniques: (i) encoder/decoder architectures; (ii) siamese neural networks; and (iii) a new family that combines the first two in hybrid architectures. These representation learning based techniques explicitly extract encodings from runtime traces before feeding them to a neural network dedicated for the end regression task on the runtime latency. We have covered deterministic and generative auto-encoders and proposed extensions of them in order to satisfy different desired encoding properties. We have also explained why the siamese neural networks are particularly interesting when a job is admitted with an arbitrary configuration and we have trained these architectures using two types of losses: a triplet loss and a soft nearest neighbor loss.

We have extended a previous benchmark of streaming workloads and sampled traces from these workloads as well as workloads from the TPCx-BB benchmark. We have provided comparative results between different modeling techniques and provided end-to-end comparative results with related work, which demonstrate the efficiency of our approach.

Our project is available online on Github: `https://github.com/udao-modeling/code`.

7.4.4 Elastic resource management in relational database systems

We have considered two problems in this setting. First, the scale-up elasticity of an online transaction processing engine (OLTP), which is required in order to allow the engine to adapt to workload variations, such as spikes in the incoming traffic, and to create new buiseness opportunities for cloud-hosted OLTP

engines. The case of OLTP elasticity is particular because transactions require a stateful concurrency control protocol to coordinate operations on the database records. Therefore, elastic scale-out approaches bring significant overheads for maintaining the protocol state, effectively reflected as additional transactional latency. We devised a novel system design which spans vertically the software virtualization stack and enables elastic scale-up of stateful engines, such as OLTP. Our results have been presented in [10].

Second, we considered the case of hybrid transactional and analytical processing engines, where a single engine needs to combine transactional (short, write-intensive queries) with analytical (long, read-only queries) workloads. We analyzed the system design space and modeled it using a set of states. We devised an adaptive algorithm which migrates across states based on the workload, and the resource availability. Our results have been presented in [21].

7.5 Explainable Anomaly Detection Benchmark

The goal of this project is to design a benchmark for precisely and reliably evaluating anomaly detection and explanation discovery techniques on high-dimensional data streams.

To do so, we first provide a dataset constituted of Spark Streaming application traces, each corresponding to the recording of a Spark Streaming application's execution on a 4-node cluster (2000+ Spark and OS-related metrics every second). This dataset contains a total of 93 traces, that recorded the runs of 10 different Spark Streaming applications. Among these 93 runs, 34 were disturbed during their recording by introducing some external events during precise periods of time, leading to range-based anomalies inside their corresponding trace. In this dataset, we consider 7 different types of events: 1) "Bursty input": abnormally high data sender's input rate, 2) "Bursty input until crash": the same event but maintained until the Spark application crashed, 3) "Stalled input": no data received from the data sender, 4) "CPU contention": abnormally high CPU usage by an external process on a given node, 5) "Driver failure": instant crash of the application's driver, 6) "Executor failure": instant crash of one of the application's executors, and 7) "Unknown": highly abnormal period observed in the collected data for an unknown reason.

To assess the performance of anomaly detection and explanation discovery techniques inside these disturbed traces, we then provide a set of metrics.

For anomaly detection and a given set of disturbed traces, techniques are evaluated using several parameter sets of "Precision and Recall for time series" [31], corresponding to 4 increasingly challenging requirements: 1) detecting the existence of anomalies 2) detecting their precise range 3) minimizing their detection latency and 4) reporting them exactly once for each ground-truth interval. For each technique, considered traces and requirements, reported metrics include the "global" PR AUC, Precision, Recall and F-score (i.e. considering all event types as the same), as well as the PR AUC and Recall per event type.

To show both the utility and usage of our benchmark for anomaly detection, we presented results for three semi-supervised techniques (respectively RNN, Autoencoder and BiGAN-based). In this setting, techniques were only tested on disturbed traces, and expected to use the undisturbed traces to model their "normal behavior". Some Spark Streaming applications being quite different from one another, metrics were reported under two modeling subjects: either training a distinct model per application, or training a single model on all applications together.

For explanation discovery, techniques are evaluated using three sets of metrics: 1) "Compactness", assessing the simplicity and understandability of the explanations for human users, 2) "Stability" and "Accuracy", assessing the "local faithfulness" of the explanations, and 3) "Conciseness" and "Generalization-ability", assessing their "global faithfulness". Once again, three techniques (MacroBase, EXstream, and LIME) were experimented with in order to show the utility of our benchmark in highlighting their strengths and weaknesses.

In 2020, we published a first report for this project available at https://arxiv.org/pdf/2010.05073.pdf.

8 Partnerships and cooperations

8.1 European initiatives

8.1.1 Collaborations with major European organizations

Angelos Anadiotis holds a visiting professor position at EPFL.

8.2 National initiatives

8.2.1 ANR

- AIDE ("A New Database Service for Interactive Exploration on Big Data") is an ANR "Young Researcher" project led by Y. Diao, started at the end of 2016.

- ContentCheck (2015-2020) is an ANR project led by I. Manolescu, in collaboration with U. Rennes 1 (F. Goasdoué), INSA Lyon (P. Lamarre), the LIMSI lab from U. Paris Sud, and the Le Monde newspaper, in particular their fact-checking team Les Décodeurs. Its aim is to investigate content management models and tools for journalistic fact-checking.

- CQFD (2019-2023) is an ANR project coordinated by F. Ulliana (U. Montpellier), in collaboration with U. Rennes 1 (F. Goasdoué), Inria Lille (P. Bourhis), Institut Mines Télécom (A. Amarilli), Inria Paris (M. Thomazo) and CNRS (M. Bienvenu). Its research aims at investigating efficient data management methods for ontology-based access to heterogeneous databases (polystores).

- SourcesSay (2020-2024) is an AI Chair funded by Agence Nationale de la Recherche and Direction Générale de l'Armement. The project goal is to interconnect data sources of any nature within digital arenas. In an arena, a dataset is stored, analyzed, enriched and connected, graph mining, machine learning, and visualization techniques, to build powerful data analysis tools.

8.2.2 Others

- The goal of the iCODA project is to develop the scientific and technological foundations for knowledge-mediated user-in-the-loop collaborative data analytics on heterogenous information sources, and to demonstrate the effectiveness of the approach in realistic, high-visibility use-cases. The project stands at the crossroad of multiple research fields—content analysis, data management, knowledge represen- tation, visualization—that span multiple Inria themes, and counts on a club of major press partners to define usage scenarios, provide data and demonstrate achievements. This is a project funded directly by Inria ("Inria Project Lab"), and is in collaboration with GraphIK, ILDA, LINKMEDIA (coordinator), as well as the press partners AFP, Le Monde (Les Décodeurs) and Ouest-France.

9 Dissemination

9.1 Promoting scientific activities

9.1.1 Scientific events: selection

Chair of conference program committees

- Ioana Manolescu was a co-chair of the demonstration track at the VLDB 2020 conference

- Ioana Manolescu was a senior PC member at the WWW 2020 conference

Member of conference program committees Ioana Manolescu was a member of the program committee for:

- CIDR (Innovative Data Systems Research) 2020
- ESWC (Extended Semantic Web Conference) 2020
- ISWC (International Semantic Web Conference) 2020
- Bases de Données Avancées (BDA) 2020

Oana Balalau has served at the following committees:

- European Chapter of the Association for Computational Linguistics (EACL 2021)
- Social Network Analysis and Mining (SNAM)
- Proceedings of the VLDB Endowment 2020 demos

A. Anadiotis has also served the PC of the VLDB Endowment 2020 demos.

Reviewer O. Balalau and P. Guzewicz have served as external reviewers at SIGMOD 2021. O. Balalau was also an external reviewer for the Symposium on Theoretical Aspects of Computer Science (STACS 2021).

9.1.2 Journal

Member of the editorial boards Ioana Manolescu has been a member of the editorial board of the Proceedings of VLDB Journal 2020.

Angelos Anadiotis has been a member of the editorial board of the Wireless Networks journal published by Springer Nature.

Reviewer - reviewing activities Angelos Anadiotis has served as reviewer for the VLDB journal.

9.1.3 Invited talks

I. Manolescu has given invited keynote talks at:

- DOLAP 2020 (22nd International Workshop On Design, Optimization, Languages and Analytical Processing of Big Data) [19]
- ADBIS 2020 (24th European Conference on Advances in Databases and Information Systems) [20]
- DATA 2020 (9th International Conference on Data Science, Technology and Applications) [24]

9.1.4 Leadership within the scientific community

Y. Diao and I. Manolescu are members of the Proceedings of Very Large Databases Endowment Board.

I. Manolescu is a member of the steering committee of the *Bases de Données Avancées* conference.

9.1.5 Scientific expertise

Oana Balalau has served as a reviewer for ANR Projects in the Artificial Intelligence Area.

9.1.6 Research administration

Ioana Manolescu is a member of the Bureau du Comité de Projets of Inria Saclay–Île-de-France. She is also responsible of the "Data Analytics and Machine Learning (DAML)" axis of the LIX, the Computer Science Laboratory of Ecole Polytechnique. The axis includes 3 teams: CEDAR, COMETE (also joint with Inria SIF) and DASCIM (LIX-only team).

Ioana Manolescu is the scientific director of LabIA, a collaboration between Inria and the DINUM focused on applying Artificial Intelligence techniques to address concrete problems encountered by public administrations. Oana Balalau is also devoting part of her activity to LabIA research projects.

9.2 Teaching - Supervision - Juries

9.2.1 Teaching

Angelos Anadiotis is full-time Assistant Professor at Ecole Polytechnique, where he is in charge of two courses:

- Master: A. Anadiotis, "Systems for Big Data", M1, Ecole Polytechnique
- Master: A. Anadiotis, "Systems for Big Data Analytics", M2, Ecole Polytechnique

I. Manolescu is a part-time (50%) professor at Ecole Polytechnique, where she is in charge of two courses:

- Master: I. Manolescu, “Database Management Systems”, 45h, M1, École Polytechnique
- Master: I. Manolescu, "Research Internship in Data Science", 30h, M1, Ecole Polytechnique

I. Manolescu is also co-responsible of the "Data Science" M1 program of Ecole Polytechnique/Institut Polytechnique de Paris (http://www.enseignement.polytechnique.fr/informatique/3A/3A.php#Data).

Team members also collaborate in teaching an M2 course of Institut Polytechnique de Paris:

- Master: A. Anadiotis, P. Guzewicz, I. Manolescu, “Architectures for Massive Data Management”, 32h, M2, Institut Polytechnique de Paris

In 2020, M. Buron and P. Guzewicz have been serving as Teaching Assistants at Ecole Polytechnique, in respectively "Data Vizualization" (taught by E. Pietriga) and "Machine Learning" (taught by J. Read).

9.2.2 Supervision

- PhD: Maxime Buron: "Raisonnement efficace sur des grands graphes hétérogènes" [25], 07/10/2020, François Goasdoué (U. Rennes 1), Ioana Manolescu and Marie-Laure Mugnier (GraphIK Inria team in Montpellier)
- PhD in progress: Ludivine Duroyon: “Data management models, algorithms & tools for fact-checking", since October 2017, François Goasdoué and Ioana Manolescu (Ludivine is in the Shaman team of U. Rennes 1 and IRISA, in Lannion)
- PhD in progress: Paweł Guzewicz: “Expressive and efficient analytics for RDF graphs”, since October 2018, Yanlei Diao and Ioana Manolescu
- PhD in progress: Qi Fan: “Multi-Objective Optimization for Data Analytics in the Cloud”, since December 2019, Yanlei Diao
- PhD in progress: Enhui Huang: “Interactive Data Exploration at Scale”, since October 2016, Yanlei Diao and Anna Liu (U. Massachussets at Amherst, USA)
- PhD in progress: Vincent Jacob: “Explainable Anomaly Detection in High-Volume Stream Analytics”, since December 2019, Yanlei Diao
- PhD in progress: Luciano di Palma, “New sampling algorithms and optimizations for interactive exploration in Big Data", since October 2017, Yanlei Diao and Anna Liu (U. Massachussets at Amherst, USA)
- PhD in progress: Khaled Zaouk: “Performance Modeling and Multi-Objective Optimization for Data Analytics in the Cloud", since October 2017, Yanlei Diao
- PhD in progress: Fanzhi Zhu: “Operator Fusion for Large Scale Complex Data Analytics Pipelines”, since July 2020, Yanlei Diao and Angelos Anadiotis

9.2.3 Juries

Angelos Anadiotis has been evaluator for PhD of Giuseppe Massimiliano Milotta who did his PhD at the Universita Mediterranea Degli Studi di Reggio Calabria.

Oana Balalau has been an examinator for PhD of Alexis Galland on ìDeep learning techniques for graph embedding at different scalesî, defended on 17 December 2020.

Ioana Manolescu has been a member of the Engineer Diploma committee of Catarina Conceiçao, at the Instituto Superior Technico, in Lisbon, Portugal, on October 16, 2020.

9.3 Popularization

9.3.1 Articles and contents

- Ioana Manolescu has been featured in a short film part of a "Women in Data Science" (WiDS) series organized as a TedX talk series and aimed at increasing the presence of female students and young professionals in Data Science (https://www.youtube.com/watch?v=KT8vmQDke4A)

9.3.2 Interventions

- Ioana Manolescu discussed fact-checking in Forum des Médias Jeunes organized par Radio Campus Lorraine, Dec 17, 2020
- Ioana Manolescu discussed accessibility and exploration of statistic data in StatsBot webinar, based on [23], next to 2020 Workshop on the Modernisation of Official Statistics organized by The United Nations Economic Commission for Europe (UNECE), on Nov 20, 2020
- Ioana Manolescu presented ContentCheck work on fact-checking at a Journée d'Etude "Répondre aux Fake News" at the University of Cergy, to an audience of faculty and students in political sciences and the media, Nov 5, 2020
- Angelos Anadiotis, Oana Balalau, Helena Galhardas, Ioana Manolescu and Youssr Youssef attended the DataHarvest Investigative Journalism Conference on Oct 24, 2020, where, at the invitation of Stéphane Horel (Le Monde), we discussed SourcesSay applications to computational journalism.

10 Scientific production

10.1 Major publications

[1] R. Alotaibi, D. Bursztyn, A. Deutsch, I. Manolescu and S. Zampetakis. 'Towards Scalable Hybrid Stores: Constraint-Based Rewriting to the Rescue'. In: *SIGMOD 2019 - ACM SIGMOD International Conference on Management of Data*. Amsterdam, Netherlands, June 2019. URL: https://hal.inria.fr/hal-02070827.

[2] M. Buron, F. Goasdoué, I. Manolescu and M.-L. Mugnier. 'Reformulation-based query answering for RDF graphs with RDFS ontologies'. In: *ESWC 2019 - European Semantic Web Conference*. Portoroz, Slovenia, Mar. 2019. URL: https://hal.archives-ouvertes.fr/hal-02051413.

[3] D. Bursztyn, F. Goasdoué and I. Manolescu. 'Teaching an RDBMS about ontological constraints'. In: *Very Large Data Bases*. New Delhi, India, Sept. 2016. URL: https://hal.inria.fr/hal-01354592.

[4] S. Cazalens, P. Lamarre, J. Leblay, I. Manolescu and X. Tannier. 'A Content Management Perspective on Fact-Checking'. In: *The Web Conference 2018 - alternate paper tracks "Journalism, Misinformation and Fact Checking"*. Lyon, France, Apr. 2018, pp. 565–574. URL: https://hal.archives-ouvertes.fr/hal-01722666.

[5] S. Cebiric, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou and M. Zneika. 'Summarizing Semantic Graphs: A Survey'. In: *The VLDB Journal* (2018). URL: https://hal.inria.fr/hal-01925496.

[6] Y. Diao, P. Guzewicz, I. Manolescu and M. Mazuran. 'Spade: A Modular Framework for Analytical Exploration of RDF Graphs'. In: *VLDB 2019 - 45th International Conference on Very Large Data Bases*. Proceedings of the VLDB Endowment, Vol. 12, No. 12. Los Angeles, United States, Aug. 2019. DOI: 10.14778/3352063.3352101. URL: https://hal.inria.fr/hal-02152844.

[7] E. Huang, L. Peng, L. D. Palma, A. Abdelkafi, A. Liu and Y. Diao. 'Optimization for active learning-based interactive database exploration'. In: *Proceedings of the VLDB Endowment (PVLDB)* 12.1 (Sept. 2018), pp. 71–84. DOI: 10.14778/3275536.3275542. URL: https://hal.inria.fr/hal-01969886.

[8] A. Roy, Y. Diao, U. Evani, A. Abhyankar, C. Howarth, R. Le Priol and T. Bloom. 'Massively Parallel Processing of Whole Genome Sequence Data: An In-Depth Performance Study'. In: *SIGMOD '17 Proceedings of the 2017 ACM International Conference on Management of Dat*. SIGMOD '17 Proceedings of the 2017 ACM International Conference on Management of Data. SIGMOD ACM Special Interest Group on Management of Data. Chicago, Illinois, United States: ACM, May 2017, pp. 187–202. DOI: 10.1145/3035918.3064048. URL: https://hal.inria.fr/hal-01683398.

10.2 Publications of the year

International journals

[9] R. Alotaibi, B. Cautis, A. Deutsch, M. Latrache, I. Manolescu and Y. Yang. 'ESTOCADA: Towards Scalable Polystore Systems'. In: *Proceedings of the VLDB Endowment (PVLDB)* 13.12 (Aug. 2020), pp. 2949–2952. DOI: 10.14778/3415478.3415516. URL: https://hal.inria.fr/hal-03150404.

[10] A.-C. Anadiotis, R. Appuswamy, A. Ailamaki, I. Bronshtein, H. Avni, D. Dominguez-Sal, S. Goikhman and E. Levy. 'A system design for elastically scaling transaction processing engines in virtualized servers'. In: *Proceedings of the VLDB Endowment (PVLDB)* 13.12 (Aug. 2020), pp. 3085–3098. DOI: 10.14778/3415478.3415536. URL: https://hal.inria.fr/hal-03104618.

[11] F. Goasdoué, P. Guzewicz and I. Manolescu. 'RDF graph summarization for first-sight structure discovery'. In: *The VLDB Journal* 29.5 (30th Apr. 2020), pp. 1191–1218. DOI: 10.1007/s00778-020-00611-y. URL: https://hal.inria.fr/hal-02530206.

International peer-reviewed conferences

[12] A. C. Anadiotis, M. Y. Haddad and I. Manolescu. 'Graph-based keyword search in heterogeneous data sources'. In: *36ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (informal publication only); 36ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (informal publication only)*. BDA 2020 - 36ème Conférence sur la Gestion de Données – Principes, Technologies et Applications. Online, France, 27th Oct. 2020. URL: https://hal.inria.fr/hal-02934277.

[13] O. Balalau and S. Goyal. 'SubRank: Subgraph Embeddings via a Subgraph Proximity Measure'. In: PAKDD 2020 - Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore, Singapore: https://pakdd2020.org/, 6th May 2020, pp. 487–498. DOI: 10.1007/978-3-030-47426-3_38. URL: https://hal.inria.fr/hal-03134181.

[14] I. Burger, I. Manolescu, E. Pietriga and F. M. Suchanek. 'Toward Visual Interactive Exploration of Heterogeneous Graphs'. In: SEAdata 2020 - Workshop on Searching, Exploring and Analyzing Heterogeneous Data in conjunction with EDBT/ICDT. Copenhagen, Denmark: https://mott.in/news/seadata-19/, 30th Mar. 2020. URL: https://hal.inria.fr/hal-02468778.

[15] M. Buron, F. Goasdoué, I. Manolescu, T. Merabti and M.-L. Mugnier. 'Revisiting RDF storage layouts for efficient query answering'. In: SSWS 2020 - 13th International Workshop on Scalable Semantic Web Knowledge Base Systems. Athène, Greece, 2nd Aug. 2020. URL: https://hal.inria.fr/hal-02921457.

[16] M. Buron, F. Goasdoué, I. Manolescu and M.-L. Mugnier. 'Obi-Wan: Ontology-Based RDF Integration of Heterogeneous Data'. In: VLDB 2020 - 46th International Conference on Very Large Data Bases. Tokyo, Japan, 31st Aug. 2020. URL: https://hal.inria.fr/hal-02921434.

[17] M. Buron, F. Goasdoué, I. Manolescu and M.-L. Mugnier. 'Ontology-Based RDF Integration of Heterogeneous Data'. In: EDBT/ICDT 2020 - 23rd International Conference on Extending Database Technology. Copenhagen, Denmark, 30th Mar. 2020. URL: https://hal.inria.fr/hal-02446427.

[18] A. Ghazimatin, O. Balalau, R. Saha and G. Weikum. 'PRINCE: Provider-side Interpretability with Counterfactual Explanations in Recommender Systems'. In: WSDM 2020 - 13th ACM International Conference on Web Search and Data Mining. Houston, Texas, United States, 3rd Feb. 2020. URL: https://hal.inria.fr/hal-02433443.

[19] I. Manolescu. 'Exploring RDF Graphs through Summarization and Analytic Query Discovery'. In: DOLAP 2020 - 22nd International Workshop On Design, Optimization, Languages and Analytical Processing of Big Data. Copenhagen, Denmark: https://www.essi.upc.edu/dtim/DOLAP/index.html, 30th Mar. 2020. URL: https://hal.inria.fr/hal-02935956.

[20] I. Manolescu. 'Integrating (Very) Heterogeneous Data Sources: A Structured and an Unstructured Perspective'. In: ADBIS 2020 - 24th European Conference on Advances in Databases and Information Systems. Lyon, France, 17th Aug. 2020, pp. 15–20. DOI: 10.1007/978-3-030-54832-2_3. URL: https://hal.inria.fr/hal-02930728.

[21] A. Raza, P. Chrysogelos, A. C. Anadiotis and A. Ailamaki. 'Adaptive HTAP through Elastic Resource Scheduling'. In: SIGMOD/PODS '20: International Conference on Management of Data. Portland OR USA, United States, 14th June 2020, pp. 2043–2054. DOI: 10.1145/3318464.3389783. URL: https://hal.inria.fr/hal-03104617.

[22] F. Song, K. Zaouk, C. Lyu, A. Sinha, Q. Fan, Y. Diao and P. Shenoy. 'Spark-based Cloud Data Analytics using Multi-Objective Optimization'. In: ICDE. Chania, Greece, 2021. URL: https://hal.inria.fr/hal-02549758.

[23] G. Thiry, I. Manolescu and L. Liberti. 'A Question Answering System For Interacting with SDMX Databases'. In: NLIWOD 2020 - 6th Natural Language Interfaces for the Web of Data / Workshop (in conjunction with ISWC). Heraklion, Greece: https://2020.nliwod.org/, 2nd Nov. 2020. URL: https://hal.inria.fr/hal-03021075.

Conferences without proceedings

[24] I. Manolescu. 'From Data to the Press: Data Management for Journalism and Fact-Checking'. In: DATA 2020 - 9th International Conference on Data Science, Technology and Applications. Paris / Virtuel, France: http://www.dataconference.org/KeynoteSpeakers.aspx#2, 7th July 2020. URL: https://hal.inria.fr/hal-02895316.

Doctoral dissertations and habilitation theses

[25] M. Buron. 'Efficient reasoning on large and heterogeneous graphs'. École Polytechnique, 7th Oct. 2020. URL: https://hal.inria.fr/tel-03107689.

Reports & preprints

[26] A. C. Anadiotis, O. Balalau, C. Conceicao, H. Galhardas, M. Y. Haddad, I. Manolescu, T. Merabti and J. You. *Graph integration of structured, semistructured and unstructured data for data journalism*. 23rd Feb. 2021. URL: https://hal.inria.fr/hal-03150441.

[27] F. Cordeiro, H. Galhardas, J. Leblay, I. Manolescu and T. Merabti. *Keyword Search in Heterogeneous Data Sources*. 30th Apr. 2020. URL: https://hal.inria.fr/hal-02559688.

[28] L. D. Palma, Y. Diao and A. Liu. *Efficient Version Space Algorithms for "Human-in-the-Loop" Model Development*. 14th Dec. 2020. URL: https://hal.inria.fr/hal-03064769.

[29] K. Zaouk, F. Song, C. Lyu and Y. Diao. *Neural-based Modeling for Performance Tuning of Spark Data Analytics*. 20th Jan. 2021. URL: https://hal.inria.fr/hal-03116831.

Other scientific publications

[30] O. Balalau, C. Conceição, H. Galhardas, I. Manolescu, T. Merabti, J. You and Y. Youssef. *Graph integration of structured, semistructured and unstructured data for data journalism*. 27th Oct. 2020. URL: https://hal.inria.fr/hal-02904797.

10.3 Cited publications

[31] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam and J. Gottschlich. *Precision and Recall for Time Series*. 2019. arXiv: 1803.03639 [cs.LG].