RESEARCH CENTRE

Inria Saclay Center at Institut Polytechnique de Paris

IN PARTNERSHIP WITH:

Institut Polytechnique de Paris

2022
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 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. – Data and knowledge analysis
- 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

- B6.5. – Information systems
- B8.5.1. – Participative democracy
- B9.5.6. – Data science
- B9.7.2. – Open data
- B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientists

- Ioana Manolescu [Team leader, INRIA, Senior Researcher, HDR]
- Oana Balalau [INRIA, ISFP]
- Oana Goga [CNRS, from Sep 2022]

Faculty Members

- Angelos Anadiotis [LIX, Associate Professor, until Apr 2022]
- Yanlei Diao [Ecole Polytechnique, Professor, HDR]

Post-Doctoral Fellows

- Madhulika Mohanty [INRIA]
- Sein Minn Sein Minn [INRIA, from Feb 2022]
- Fei Song [Ecole Polytechnique]
- Prajna Upadhyay [Inria, until Sep 2022]

PhD Students

- Salim Chouaki [CNRS]
- Nelly Barret [INRIA]
- Theo Bouganim [INRIA]
- Tom Calamai [Amundi, CIFRE]
- Qi Fan [LIX]
- Mhd Yamen Haddad [INRIA]
- Vincent Jacob [LIX]
- Muhammad Khan [INRIA]
- Kun Zhang [INRIA]

Technical Staff

- Arnab Sinha [Ecole Polytechnique, Engineer]
- Simon Ebel [INRIA, Engineer, from Mar 2022]
- Theo Galizzi [INRIA, Engineer, from Apr 2022]
- Joffrey Thomas [Ecole Polytechnique, Engineer]

Interns and Apprentices

- Louis Caubet [Ecole Polytechnique, from Oct 2022]
- Antoine Gauquier [Inria, Intern, from May 2022 until Sep 2022]
- Quentin Massonnat [Ecole Polytechnique, Intern, from Apr 2022 until Jul 2022]
- Nina Varchavsky-Bergin [Inria, Intern, from Jun 2022 until Aug 2022]

Administrative Assistant

- Maria Ronco [INRIA]

External Collaborators

- Rana Alotaibi [UNIV CALIFORNIE]
- Mathilde Bouquerel [Radio France, from Nov 2022]
- Estelle Cognacq [Radio France, from Aug 2022]
- Antoine Deiana [Radio France, from May 2022]
- Emilie Gautreau [Radio France, from May 2022]
- Stéphane Horel [LE MONDE]
- Antoine Krempf [Radio France, from Apr 2022]
- Chenghao Lyu [University of Massachusetts, Amherst]
- Adrien Maumy [Radio France, from May 2022]
- Thomas Pontillon [Radio France, from May 2022]
- Gerald Roux [Radio France, from May 2022]
- Saumya Yashmohini Sahai [OSU]
- Joanna Yakin [Radio France, from May 2022]

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. In addition, we explore and mine rich data via machine learning techniques. Our scientific contributions fall into four interconnected areas:

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-objective optimization are leveraged to build performance models for cloud data analytics. The same goal is shared by our work on the efficient evaluation of queries in dynamic knowledge bases.

Data discovery and exploration. Today's Big Data is complex; understanding and exploiting it is daunting, especially to novice users such as journalists or domain scientists. To help such users, in the AI Chair "SourcesSay: Intelligent Data Analysis and Interconnection in Digital Arenas", we explore efficient and keyword search techniques to find answers in the data its highly heterogeneous structure makes standard (e.g., SQL) queries inapplicable. Further, we propose novel data abstraction methods, which, given a dataset, automatically compute a simple, human-understandable model thereof. Finally, we study heterogeneous graph exploration, blending graph querying, and natural language summarization.

Natural language understanding for analyzing and supporting digital arenas. In this area, we are focused on new NLP tools and their applications to problems such as argumentation mining, information extraction, and question answering. Natural language is ubiquitous and better tools for extracting, classifying, and generating text are needed for various applications.

Safeguarding information systems. Recent events have brought to light the easiness of using current online systems to propagate information (that is sometimes false) and that we are facing an information war. We create knowledge and technology in this area to make the online information space safer. In O. Goga's ERC project "Momentous: Measuring and Mitigating Risks of AI-driven Information Targeting", we seek to use AI for good to help fact-checking and journalists, we develop natural language processing techniques to detect malicious online content (e.g., propaganda, manipulation), and we develop measurement methodologies and controlled experiments to assess risks with online systems.

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*. However, 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 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 unstructured databases and social feeds. However, current content management tools for this task are not suited for the task, in particular when they require a lengthy rigid cycle of data integration and consolidation in a warehouse. Thus, we need flexible tools allowing us to interconnect various kinds of data sources and query them together.

3.3 Natural language question answering

We investigate methods for finding useful information in large datasets to provide support for investigative journalism and not only. For example, real-world events such as elections, public demonstrations, disclosures of illegal or surprising activities, etc., are mirrored in new data items being created and added to the global corpus of available information. Making sense of this wealth of data by providing a natural language question-answering framework will facilitate the work of journalists, but it can also be extremely useful to non-technical users in general.

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 for 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 " explore-by-example " approach.

3.5 Exploratory querying of data graphs

Semantic graphs, including data and knowledge, are hard to apprehend for users due to the complexity of their structure and, often to their large volumes. To help tame this complexity, our research follows several avenues. First, we build compact summaries of Semantic Web (RDF) graphs suited for a first-sight interaction with the data. Second, we devise fully automated methods of exploring RDF graphs using interesting aggregate queries, which, when evaluated over a given input graph, yield interesting results (with interestingness understood in a formal, statistical sense). Third, we study the exploration of highly heterogeneous data graphs resulting from integrating structured, semi-structured, and unstructured (text) data. In this context, we develop data abstraction methods, showing the structure of any dataset to a novice user, as well as searching on the graph through (i) keyword queries and (ii) exploration leveraging graph structure and linguistic contents.

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 develop a data analytics optimizer that can automatically determine a cluster configuration with a suitable number of cores and 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 to leverage the opportunities for efficient resource management by adapting to the variations and heterogeneity of the workloads. Resource management in a virtualized setting, like the cloud, must be enforced in a performance-efficient manner to avoid introducing overheads to the execution. We design elastic systems that change their configuration at runtime with minimal cost to adapt to the workload every time. Changes in the design include both different resource allocations and different data layouts. We consider different workloads, including transactional, analytical, and mixed, and we study the performance implications on different configurations to propose a set of adaptive algorithms.

3.8 Argumentation mining

Argumentation appears when we evaluate the validity of new ideas, convince an addressee, or solve a difference of opinion. An argument contains a statement to be validated (a proposition also called claim or conclusion), a set of backing propositions (called premises, which should be accepted ideas), and a logical connection between all the pieces of information presented that allows the inference of the conclusion. In our work, we focus on fallacious arguments, where evidence does not prove or disprove the claim, for example, in an "ad hominem" argument, a claim is declared false because the person making it has a character flaw. We study the impact of fallacies in online discussions and show the need for improving tools for their detection. In addition, we look into detecting verifiable claims made by politicians. We started a collaboration with RadioFrance and with Wikidébats, a debate platform focused on proving quality arguments for controversial topics.

3.9 Measuring and mitigating risks of AI-driven information targeting

We are witnessing a massive shift in the way people consume information. In the past, people had an active role in selecting the news they read. More recently, the information started to appear on people's social media feeds as a byproduct of one's social relations. We see a new shift brought by the emergence of online advertising platforms where third parties can pay ad platforms to show specific information to particular groups of people through paid targeted ads. AI-driven algorithms power these targeting technologies. Our goal is to study the risks with AI-driven information targeting at three levels: (1) human-level–in which conditions targeted information can influence an individual's beliefs; (2) algorithmic-level–in which conditions AI-driven targeting algorithms can exploit people's vulnerabilities; and (3) platform- level–are targeting technologies leading to biases in the quality of information different groups of people receive and assimilate. Then, we will use this understanding to propose protection mechanisms for platforms, regulators, and users.

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 the journalists from RadioFrance, the team Le vrai du faux.

4.3 Computational social science

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 investigate the impact of propaganda in online forums and we study a particular type of propagandist content, the fallacious argument. We show that identifying such arguments remains a difficult task, but one of high importance because of the pervasiveness of this type of discourse. We also explore trends around the diffusion and consumption of propaganda and how this can impact or be a reflection of society.

4.4 Online targeted advertising

The enormous financial success of online advertising platforms is partially due to the precise targeting features they offer. Ad platforms collect large amounts of data on users and use powerful AI-driven algorithms to infer users' fine-grain interests and demographics, which they make available to advertisers to target users. For instance, advertisers can target groups of users as small as tens or hundreds and as specific as "people interested in anti-abortion movements that have a particular education level". Ad platforms also employ AI-driven targeting algorithms to predict how "relevant" ads are to particular groups of people to decide to whom to deliver them. While these targeting technologies are creating opportunities for businesses to reach interested parties and lead to economic growth, they also open the way for interested groups to use user's data to manipulate them by targeting messages that resonate with each user.

5 Social and environmental responsibility

5.1 Impact of research results

Our work on Big Data and AI techniques applied to data journalism and fact-checking have attracted attention beyond our community and was disseminated in general-audience settings, for instance through I. Manolescu's participation in panels at *Médias en Seine*, at the Colloque Morgenstern at Inria Sophia, and through invited keynotes, e.g., at DEBS 2022 and DASFAA 2022.

Our work in the SourcesSay project (Section 8.1.1), on propaganda detection (Section 8.2.1), and on ad transparency (Section 8.5), goes towards making information sharing on the Web more transparent and more trustworthy.

6 Highlights of the year

6.1 Awards

Quentin Massonnat (M1 intern, Ecole Polytechnique, advised by O.Balalau and I. Manolescu) received for his M1 thesis the Prix du Centre de Recherche from Ecole Polytechnique.

6.2 Collaboration with RadioFrance

The team has started a collaboration with RadioFrance, the national radio operator, developing a new tool for automatically detecting and verifying (when possible) statistic and other claims. The tool has been made available to journalists who already use it, and it has lead to several international publications [14, 13]. The team is grateful for the support provided by Inria and our research center towards our collaboration with RadioFrance.

The team finds important to thank the outstanding efforts made by the Inria Commission d'Evaluation, organizing and participating to Inria hiring and promotion committees, keeping us, researchers, meticulously informed, and upholding the moral and intellectual values we are collectively proud of, and which define our institute.

7 New software and platforms

7.1 New software

7.1.1 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-Denisa Balalau, Catarina Pinto Conceicao

Contact: Manolescu Ioana

7.1.2 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

7.1.3 Butterfly

Keywords: Data management, Databases, Graph processing

Functional Description: Integrated system for data science workload processing. Butterfly includes operators for relational and graph processing, as well as different data layouts and execution models.

Contact: Angelos Anadiotis

7.1.4 GraphCite

Name: GraphCite: Citation Intent Classification in Scientific Publications via Graph Embeddings

Keywords: Graph embedding, Citation intent prediction

Functional Description: Citations are crucial in scientific works as they help position a new publication. Each citation carries a particular intent, for example, to highlight the importance of a problem or to compare against results provided by another method. The authors' intent when making a new citation has been studied to understand the evolution of a field over time or to make recommendations for further citations. In this software, we address the task of citation intent prediction from a new perspective. In addition to textual clues present in the citation phrase, we also consider the citation graph, leveraging high-level information of citation patterns.

Contact: Oana-Denisa Balalau

7.1.5 Abstra

Name: Abstra: Toward Generic Abstractions for Data of Any Model

Keywords: Heterogeneous Data, Data Exploration, Data analysis, Databases, LOD - Linked open data

Functional Description: Abstra computes a description meant for humans, based on the idea that, regardless of the syntax or the data model, any dataset holds some collections of entities/records, that are possibly linked with relationships. Abstra relies on a common graph representation of any incoming dataset, it leverages Information Extraction to detect what the dataset is about, and relies on an original algorithm for selecting the core entity collections and their relations. Abstractions are shown both as HTML text and a lightweight Entity-Relationship diagram. A GUI also allows to tune the abstraction parameters and explore the dataset.

URL: `https://team.inria.fr/cedar/projects/abstra/`

Contact: Nelly Barret

7.1.6 StatCheck

Name: Fact-checking Multidimensional Statistic Claims in French

Keywords: Machine learning, Databases, Natural language processing, Software engineering

Scientific Description: To strengthen public trust and counter disinformation, computational fact-checking, leveraging digital data sources, attracts interest from the journalists and the computer science community. A particular class of interesting data sources comprises statistics, that is, numerical data compiled mostly by governments, administrations, and international organizations. Statistics are often multidimensional datasets, where multiple dimensions characterize one value, and the dimensions may be organized in hierarchies. This paper describes STATCHECK, a statistic fact-checking system jointly developed by the authors, which are either computer science researchers or fact-checking journalists working for a French-language media with a daily audience of more than 15 millions (aud, 2022). The technical novelty of STATCHECK is twofold: (i) we focus on multidimensional, complex-structure statistics, which have received little attention so far, despite their practical importance, and (ii) novel statistical claim extraction modules for French, an area where few resources exist. We validate the efficiency and quality of our system on large statistic datasets (hundreds of millions of facts), including the complete INSEE (French) and Eurostat (European Union) datasets, as well as French presidential election debates.

Functional Description: StatCheck firstly allows the collection of data for its operation. Two types of data are collected: statistical tables and posts from social networks: - Acquisition of statistical files on the site of referent organisations (INSEE, Eurostat) - Extraction of statistical tables from these files, and storage of the extracted tables - Acquisition of political tweets from a list of accounts The application allows the detection, extraction and search of statistical facts: - Detection and extraction of statistical facts from Twitter posts (e.g. "Unemployment rate increased by 30% in 2023) - Search for statistical facts in our database. Display of the twenty most relevant statistical tables for a statistical fact - Automatic transcription of audio files to detect and extract transcripts of statistical facts.

Release Contributions: - Redesign of the user interface - Modification of the software architecture - Addition of audio transcription

URL: `https://cedar-rf.saclay.inria.fr/`

Publications: hal-01496700, hal-01745768, hal-02121389, hal-01915148, hal-03767992, hal-03791175

Contact: Ioana Manolescu

Participants: Tien Duc Cao, Ioana Manolescu, Xavier Tannier, Oana-Denisa Balalau, Simon Ebel, Theo Galizzi

8 New results

8.1 Data management for analyzing digital arenas

8.1.1 Graph integration of heterogeneous data sources for data journalism

Work carried within the ANR AI Chair SourcesSay project 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 [11]. The main technical challenges are: (i) how to interconnect structured and semi-structured data sources? We address this through 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 specific search criteria, or certain keywords? The question is particularly challenging in our context since ConnectionLens graphs can be pretty large, and query answers can traverse edges in both directions.

In this context, the following new contributions have been brought:

1. **Integrating connection search in graph queries.** When graphs are very heterogeneous and/or users are unfamiliar with their structure, they may need to find how two or more groups of nodes are connected in a graph, even when users cannot describe the connections. This is only partially supported by existing query languages, which allow searching for paths, but not for trees connecting three or more node groups. Prior work on keyword search in databases tackled variants of this problem, related, in its most general form, to the NP-hard Group Steiner Tree problem. We show how to integrate connecting tree patterns (CTPs, in short) with a graph query language such as GPML, SPARQL, or Cypher, leading to Extended Queries. We then study a set of algorithms for evaluating CTPs; we generalize prior keyword search work to be complete, most importantly by (i) considering bidirectional edge traversal, (ii) allowing users to select any score function for ranking CTP results, and (iii) returning all results. Finally, to cope with very large search spaces, we developed efficient pruning techniques and formally established a large set of cases where our best algorithm, MOLESP, is complete even with pruning. Its benefits are validated on synthetic and real-world workloads [21, 27].

2. **Shared GAM: generalizing GAM search over graphs for batch keyword query evaluation.** GAM is a novel query algorithm for keyword search in graphs. Given a query as a set of search terms, GAM finds links between them within the graph. However, when multiple keyword queries share some of their keywords, GAM runs once per query, leading to duplicate computations for the common keywords and, therefore, inefficiency. In this work, we aim to eliminate this inefficiency by developing a shared version of GAM (we refer to it as SharedGAM) algorithm to execute such keyword queries together by sharing computations.

3. **Abstra: toward generic abstractions for data of any model.** Abstra is an all-in-one dataset abstraction system that produces an Entity-Relationship schema and a textual description out of any (semi)-structured dataset (e.g., XML documents, JSON documents, RDF graphs, property graphs). Such a system helps users to grasp the content of a dataset they don't know and decide whether it is useful for their needs. Moreover, many actors might benefit from such abstractions, such as scientific researchers, (data) journalists, data providers, and consumers. In a nutshell, Abstra starts by transforming any incoming dataset into a graph. Then, it summarizes it and detects the main entities and their relationships. Main entities are further classified into user-friendly categories. Finally, the abstraction result is presented in the form of an HTML page, showing the textual description of the dataset and the Entity-Relationship schema, both built with the main entities and their relationships found previously. This work has led to a demonstration paper [15, 25].

4. **Identifying disguised missing values in heterogeneous, text-rich data** Digital data is produced in many data models, ranging from highly structured (typically relational) to semi-structured models (XML, JSON) to various graph formats (RDF, property graphs) or text. Most real-world datasets contain a certain amount of null values, denoting missing, unknown, or inapplicable information. While some data models allow representing nulls by special tokens, so-called disguised missing values (DMVs, in short) are also frequently encountered: these are values that are not syntactically speaking nulls but which do, nevertheless, denote the absence, unavailability, or inapplicability of the information. In this work, we tackle the detection of a particular kind of DMV: texts freely entered by human users. This problem is not tackled by DMV detection methods focused on numeric or categoric data; further, it also escapes DMV detection methods based on value frequency since such free texts are often different from each other, thus most DMVs are unique. We encountered this problem within the ConnectionLens project, where heterogeneous data is integrated into large graphs. We presented two DMV detection methods for our specific problem: (i) leveraging Information Extraction, already applied in ConnectionLens graphs; and (ii) through text embeddings and classification. We detail their performance-precision trade-offs on real-world datasets [26].

ConnectionLens is available online at: https://gitlab.inria.fr/cedar/connection-lens.

8.1.2 Fact-checking Multidimensional Statistic Claims in French

To strengthen public trust and counter disinformation, computational fact-checking, leveraging digital data sources, attracts interest from journalists and the computer science community. A particular class of interesting data sources comprises statistics, that is, numerical data compiled mostly by governments, administrations, and international organizations. Statistics are often multidimensional datasets, where multiple dimensions characterize one value and the dimensions may be organized in hierarchies. To address this challenge we developed STATCHECK, a statistic fact-checking system, in collaboration with RadioFrance. The technical novelty of STATCHECK is twofold: (i) we focus on multidimensional, complex-structure statistics, which have received little attention so far, despite their practical importance; and (ii) novel statistical claim extraction modules for French, an area where few resources exist. We validate the efficiency and quality of our system on large statistic datasets (hundreds of millions of facts), including the complete INSEE (French) and Eurostat (European Union) datasets, as well as French presidential election debates [13, 14].

8.2 Natural language understanding for analyzing and supporting digital arenas

8.2.1 Argumentation mining

Humans use argumentation daily to evaluate the validity of new ideas, convince an addressee, or solve a difference of opinion. An argument contains a statement to be validated (a proposition also called claim or conclusion), a set of backing propositions (called premises, which should be accepted ideas), and a logical connection between all the pieces of information presented that allows the inference of the conclusion. In this work, we will focus on **fallacies**: weak arguments that seem convincing, however, their evidence does not prove or disprove the argument's conclusion. Fallacy detection is part of argumentation mining, the area of natural language processing dedicated to extracting, summarizing, and reasoning over human arguments. The task is closely related to propaganda detection, where propaganda consists of a set of manipulative techniques, such as fallacies, used in a political context to enforce an agenda [2]. In the past, we have worked on propaganda [2] and fallacy detection [10]. We continue this work with a CIFRE PhD that started this year, a collaboration between the Amundi company, Inria and Télécom Paris. This thesis aims to improve fallacy detection in natural language by leveraging both language patterns and additional information, such as common sense knowledge, encyclopedic knowledge and logical rules. To achieve this we will focus on how fallacies can be represented and how we can classify reasoning patterns in argumentation. The interest of Amundi is in how argumentation can be applied to finding examples of greenwashing.

8.2.2 Finding conflicts of opinion on citizen participation platforms

Online citizen participation platforms allow large numbers of contributors to be involved in public decision-making, overcoming limiting factors for their offline counterparts, such as their geographical position. However, for large groups of contributors to collaborate and co-construct joint, well-elaborated proposals, we need to provide tools for users and decision-makers to navigate and understand high volumes of content. To achieve this, we introduce an approach based on natural language processing to detect pairs of contradictory and equivalent proposals in online citizen participation contexts. We apply this approach to two major national citizen consultations, namely the *République Numérique* and *Revenu Universel d'Activité* consultations. Our method leverages a Transformer-based classifier, fine-tuned on natural language inference datasets and on two weakly labeled datasets created using data from these consultations. We also address the classification problem on large texts by proposing alternative strategies explicitly designed for texts containing more than one sentence. Finally, we highlight the great potential of our tool in the analysis, synthesis, and recommendation of contributions to citizen participation platforms. This work is currently submitted for publication.

8.2.3 Open information extraction with entity-focused constraints

Open Information Extraction (OIE) is the task of extracting tuples of the form (subject, predicate, object), without any knowledge of the type and lexical form of the predicate, the subject, or the object. In this work, we focus on improving OIE quality by exploiting domain knowledge about the subject and object. More precisely, knowing that the subjects and objects in sentences are oftentimes named entities, we explore how to inject constraints in the extraction through constrained inference and constraint-aware training. Our work leverages the state-of-the-art OpenIE6 platform, which we adapt to our setting. Through a carefully constructed training dataset and constrained training, we obtain a 29.17% F1-score improvement in the CaRB metric and a 24.37% F1-score improvement in the WIRe57 metric. Our technique has important applications, specifically in the construction of vertical knowledge bases supporting investigative journalism, where automatically extracting conflict-of-interest between scientists and funding organizations helps understand the type of relations companies engage with the scientists. This work is under submission.

8.2.4 Citation intent prediction in scientific articles

Citations are crucial in scientific works as they help position a new publication. Each citation carries a particular intent, for example, to highlight the importance of a problem or to compare against results provided by another method. The authors' intent when making a new citation has been studied to understand the evolution of a field over time or to make recommendations for further citations. In this work, we address the task of citation intent prediction from a new perspective. In addition to textual clues present in the citation phrase, we also consider the citation graph, leveraging high-level information of citation patterns. In this novel setting, we perform a thorough experimental evaluation of graph-based models for intent prediction. We show that our model, GraphCite, improves significantly upon models that consider only the citation phrase [16].

8.2.5 Graph-to-text for question generation

The first topic we explored in our work is **automatically generating suitable questions for existing knowledge bases**. A knowledge graph (KB) is represented by a set of triples, where each triple is composed of a subject, predicate, and object. Our question generation model aims to generate a set of answerable questions from a given knowledge graph KB, where each question corresponds to a subgraph of KB. We are investigating training Transformer networks to generate questions from knowledge graphs. This approach is based on existing datasets that match questions to subgraphs (e.g., SimpleQuestions, GraphQuestions, GrailQA, etc.). The subgraphs in these datasets are usually from existing KBs such as FreeBase and WikiData. An important challenge of this approach is that the neural network usually fails to predict the correct sentences under zero-shot settings, i.e. when it encounters unseen predicates or entities in the test set. This problem is especially acute in the case of zero-shot predicates. We intend to use unsupervised methods for the neural network to learn to understand the unseen predicates and entities to help the question generation under zero-shot settings.

In parallel, we are aligning Question-Answering datasets across similar KBs. We hope that this work will help us have larger training sets for graph-to-text generation of questions. At the core, this problem involves aligning Freebase's entities, classes, and predicates to those of YAGO4. This is especially interesting because (i) even though YAGO has been around for quite some time, there is a dearth of QA datasets on it, and (ii) there are many QA datasets on Freebase, but Google is no longer maintaining Freebase. We have used the paraphrase model of BERT for computing predicate matching with some success. Inspired by previous works, we have also developed a Greedy Matching algorithm for iteratively aligning the two KGs. Going forward, we will now evaluate the Bert model's performance, improve the Greedy Matcher's results, and finally generate a QA dataset on YAGO4.

8.3 Efficient Big Data analytics

8.3.1 Towards speeding Up graph-Relational queries in RDBMSs

Graph data is generally stored and processed using two main approaches: (i) extending existing relational database management systems (RDBMSs) with graph capabilities, and (ii) through native graph database management systems (GDBMSs). The advantage of leveraging RDBMSs is to benefit from the maturity of their query optimization and execution. Conversely, native GDBMSs treat complex graph structures as a first-class citizens, which may make them more efficient on complex structural queries. In this work, we consider the processing of graph-relational queries, that is, queries mixing graph and relational operators, on graph data. We take a purely relational approach, reorganizing the graph connectivity information using a novel CSR Optimised Schema (COS). Based on our storage model, incoming queries are reformulated to account for the COS data organization, which can then be optimized and executed by an RDBMS. We have implemented our approach on top of PostgreSQL and we demonstrate that COS improves the performance for many graph-relational queries of the popular Social Network Benchmark [23].

8.3.2 Scalable analytics on multi-streams dynamic graphs summary

Several real-time applications rely on dynamic graphs to model and store data arriving from multiple streams. In addition to the high ingestion rate, the storage and query execution challenges are amplified in contexts where consistency should be considered when storing and querying the data. This Ph.D. thesis addresses the challenges associated with multi-stream dynamic graph analytics. We propose a database design that can provide scalable storage and indexing to support consistent read-only analytical queries (present and historical), in the presence of real-time dynamic graph updates that arrive continuously from multiple streams [22].

8.3.3 Fine-grained modeling and optimization for intelligent resource management in Big Data processing

Big data processing at the production scale presents a highly complex environment for resource optimization (RO), a problem crucial for meeting analytical users' performance goals and budgetary constraints. The RO problem is challenging because it involves a set of decisions (the partition count, placement of parallel instances on machines, and resource allocation to each instance), requires multi-objective optimization (MOO), and is compounded by the scale and complexity of big data systems while having to meet stringent time constraints for scheduling. This project addressed the resource optimization problem for a custom-built big data processing system (MaxCompute) of the Alibaba Cloud. It supports multi-objective resource optimization via fine-grained instance-level modeling and optimization. We proposed a new architecture that breaks RO into simpler problems, new fine-grained predictive models, and novel optimization methods that exploit these models to make effective instance-level RO decisions well under a second [18].

8.4 Anomaly detection

8.4.1 Fast and Explainable Anomaly Detection over High-Dimensional Data Streams

In our increasingly digital and connected society, high-volume data streams have become more and more prevalent and complex. This has made incidents more likely, diverse, and therefore harder to manually anticipate and diagnose for humans. In this project, we aim to assist such anticipation and diagnosis through automated solutions, focusing on deep, unsupervised, and explainable anomaly detection. We conducted multiple comparative studies on the recently proposed Exathlon benchmark, focusing on two main use cases. In the first use case, we aim to detect and explain anomalies in time series recordings from repeated executions of Spark Streaming applications. In this use case, we observed that reconstruction-based and forecasting-based detection methods could generalize well to different execution environments while exhibiting complementary behaviors. In the second use case, we target anomalies in financial transactions coming from the Swift messaging network. In both use cases, our current research focuses on designing new methods that improve detection accuracy and explanation consistency, with a direction being to leverage a few anomaly labels, for instance via contrastive learning approaches.

8.4.2 Accurate, explainable anomaly detection through visually-informative projections

Despite their widespread adoption, anomaly detection (AD) systems thus far have mainly focused on detection power. However, emerging applications such as "Artificial Intelligence for IT operations" (AIOps) point to the need for "explainable anomaly detection" to enhance business operations with proactive, personalized, and dynamic insight and further enable corrective or preventive action to resolve IT performance issues. In our ongoing work, we address explainable AD by proposing a new explainable AD model, which achieves explainability via a set of visually-informative patterns in low-dimensional axis-aligned projections while retaining prediction accuracy. Our model, called VIPAD, builds on a classical explainable classification framework, VIPR, but addresses its fundamental limitations for anomaly detection. Our evaluation using the latest anomaly detection benchmark Exathlon for AIOps shows that VIPAD can approximate the accuracy of random forests, which is not explainable while outperforming other explainable models in both prediction accuracy and quality of explanations.

8.5 Collaborative ad transparency: promises and limitations

Several targeted advertising platforms offer transparency mechanisms, but researchers and civil societies repeatedly showed that those have major limitations. In this paper, we propose a collaborative ad transparency method to infer, without the cooperation of ad platforms, the targeting parameters used by advertisers to target their ads. Our idea is to ask users to donate data about their attributes and the ads they receive and to use this data to infer the targeting attributes of an ad campaign. We propose a Maximum Likelihood Estimator based on a simplified Bernoulli ad delivery model. We first test our inference method through controlled ad experiments on Facebook. Then, to further investigate the potential and limitations of collaborative ad transparency, we propose a simulation framework that allows varying key parameters. We validate that our framework gives accuracies consistent with real-world observations such that the insights from our simulations are transferable to the real world. We then perform an extensive simulation study for ad campaigns that target a combination of two attributes. Our results show that we can obtain good accuracy whenever at least ten monitored users receive an ad. This usually requires a few thousand monitored users, regardless of population size. Our simulation framework is based on a new method to generate a synthetic population with statistical properties resembling the actual population, which may be of independent interest.

9 Partnerships and cooperations

9.1 European initiatives

9.1.1 Horizon Europe

- Oana Goga is the PI – ERC Starting Grant 2022 – 2027 “MOMENTOUS: Measuring and Mitigating Risks of AI-driven Information Targeting” (1,499,952 €).

9.1.2 H2020 projects

- Oana Goga is the local PI for CNRS partner – EU H2020 2021 – 2024 “Trust aWARE: Enhancing Digital Security, Privacy and TRUST in softWARE” (our part: 461,000 €).

9.2 National initiatives

9.2.1 ANR

- Oana Goga is the local PI for LIX partner – ANR PRC 2022 – 2026 “FeedingBias: A multi-platform mixed-methods approach to news exposure on social media” (our part: 128,000 €)

- Oana Goga is the local PI for LIX partner – ANR PRCE 2021 – 2025 “PROPEOS: Privacy-oriented Personalization of Online Services” (our part: 202,720 €)

- CQFD (2019-2024) 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.

9.2.2 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.

- ANRT Project: CIFRE Amundi, advised by O.Balalau and F.Suchanek (Télécom Paris). The goal of this thesis is to improve fallacy detection in natural language, by leveraging both language patterns but also additional information, such as common sense knowledge, encyclopedic knowledge and logical rules. To achieve this we will focus on how fallacies can be represented1 and how we can classify reasoning patterns in argumentation.

9.3 Regional initiatives

- Hi!Paris Collaborative Project (2022-2024) coordinated by O.Balalau and J.Romero (Télécom Sud-Paris). The project will be able to improve interactions between humans and machines by better extracting structured information from natural language.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: selection

Chair of conference program committees

- Yanlei Diao: Chair of the 2022 ACM SIGMOD Awards Committee

- Ioana Manolescu: "Social Web" Track chair at the Web Conference 2022

Member of the conference program committees

- Oana Balalau: EMNLP 2022 (The 2022 Conference on Empirical Methods in Natural Language Processing)

- Oana Goga: The Web Conference (2023), USENIX Security (2023), CoNEXT Student Workshop (2022)

- Ioana Manolescu: CIDR (Conference on Innovative Database Systems) 2022, EDBT (Extending Database Technology) 2022, AI/ML 2022

- Madhulika Mohanty: CoDS-COMAD 2022, SIGMOD (Demo) 2022, SIGMOD Availability 2021-2022, TheWebConf (WWW) 2022, SIGIR (Demo) 2022, AIMLSystems 2022

- Sein Minn: AIED (AI in Education) 2022, IEEE BigData 2022

- Prajna Upadhyay : ACM SIGMOD 2022 Availability & Reproducibility, AIML Systems 2022, ICONIP 2022

Reviewer

- Oana Balalau: TheWebConf 2022
- Madhulika Mohanty: CIDR 2022
- Prajna Upadhyay: EMNLP 2022, CIKM 2022 Short papers, KDD 2022 Applied Data Science track, EDBT 2022, CIDR 2022

10.1.2 Journal

Member of the editorial boards

- Ioana Manolescu: Associate Editor for PVLDB 2022

Reviewer - reviewing activities

- Oana Balalau: Information Processing Letters
- Sein Minn: Springer Journal of Data Mining and Knowledge Discovery, MDPI Journal of Information, IEEE Transactions on Education, IEEE Transactions on Learning Technologies
- Arnab Sinha: IEEE Internet of Things Journal

10.1.3 Invited talks

- Angelos Anadiotis presented teamwork on conflicts of interest in the biomedical domain at the University of Cornell Database Seminar.
- Oana Balalau:
  - Keynote "Moving Towards Better Online Argumentation" in workshop Deep Learning pour le traitement automatique des langues, associated with EGC 2022
  - Seminar NoRDF at Télécom Paris on "Argumentation Mining: Challenges and Opportunities in Detecting Fallacious Reasoning."
- Yanlei Diao
  - Distinguished Lecture, Berlin Institute for the Foundations of Learning and Data, September 15, 2022
- Oana Goga:
  - Keynote "Can we safeguard the micro-targeting of political ads? An algorithmic and regulatory perspective" at Colloque Big Data ISTC, Lille, October 2022
  - Seminar at INSA Lyon (DRIM@LIRIS Seminar), Lyon, October 2022
- Ioana Manolescu
  - "Teasing Journalistic Findings out of Heterogeneous Sources: A Data/AI journey" as a keynote at the DASFAA 2022 conference, then in the Colloque Jacques Morgenstern at Sophia Antipolis, on June 2, 2022, then on June 28 at the DEBS 2022 conference.
  - "Que disent les sources ? L'IA et le BigData au service de la détection des fausses nouvelles", seminar for the Digital section of Académie des Technologies
  - Participation to the panel "Quand la tech se met au service de l'information" at the "Médias en Seine" festival, on Nov 22, 2022

10.1.4 Leadership within the scientific community

- Oana Goga: Co-leader of the Action PLATFORM in GDR CIS.

10.1.5 Scientific expertise

- Oana Balalau: reviewed grant application for Viena Science Fund
- Oana Goga:
  - Advisory Boards: Member of the Science Advisory Committee for the NSF-funded Mid-scale RI-1 project (a 15 million US $ project): Observatory for online human and platform behavior (2022).
  - Institute Evaluation Committee: Evaluation of the Max Planck Graduate Center Computer Science (Jan 2023)
  - Award Committee: CNIL-Inria Privacy Protection Award (2022)
  - Grant reviewing: ERC Starting Grant (2022), German National Research Center for Applied Cybersecurity ATHEN (2022).

10.1.6 Research administration

- Ioana Manolescu
  - Joined the Comité Operationnel of Hi!Paris, an Institut Polytechnique de Paris organization fostering research on this topic, in October 2022.
  - Been a member of Inria's Bureau des Comités de Projet from January to October 2022.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

Until April 2022, A. Anadiotis has been a full-time Assistant Professor at Ecole Polytechnique, where he has been 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

O. Balalau is a part-time (33%) assistant professor at Ecole Polytechnique, where she teaches two courses:

- INF473G “Graphe Global Géant”, L3, Ecole Polytechnique
- INF583 “Systems for Big Data”, M1, Ecole Polytechnique

Yanlei Diao: University of Massachusetts Amherst, CMPSCI645, January 24 - March 11, 2022.

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

- INF553 (Database Management Systems), M1, Ecole Polytechnique, 45 ETD;
- INF592 (Internships in Artificial Intelligence and Data Science), together with Jesse Read;
- "Artificial Intelligence and Data Science" year of study (3rd year for Polytechnique students).

I. Manolescu also teaches a course at Institut Mines Télécom:

- TPT DATAIA921 (Architectures for Big Data Management), M2, 18 ETD

Team members also collaborate in teaching courses at Institut Polytechnique de Paris:

- Nelly Barret:
  - labs of INF371 (Mécanismes de la programmation orientée-objet)
  - labs of INF411 (Les bases de la programmation et de l'algorithmique)
- Madhulika Mohanty: TPT DATAIA921 (Architectures for Big Data Management), M2, Institut Mines Télécom, 9h TP.
- Y. Hadad has been a teaching assistant in INF553 Database Management Systems, taught by I. Manolescu.

10.2.2 Supervision

PhD supervision:

- N. Barret, from January to December 2022 (I. Manolescu)
- T. Bouganim from October to December 2022(I. Manolescu, E. Pietriga)
- T. Calamai, from December 2022 (O. Balalau, F.Suchanek)
- S. Chouaki, from December 2022 (O. Goga)
- Qi Fan, from January to December 2022 (Y. Diao)
- Y. Haddad, from January to December 2022 (A. Anadiotis, I. Manolescu)
- G. Khan, from January to December 2022 (A. Anadiotis, I. Manolescu)
- V. Jacob, from January to December 2022 (Y. Diao)
- C. Lyu, from January to December 2022 (Y. Diao, visiting PhD From UMass Amherst)
- V. Sosnovik, from September to December 2022 (O. Goga)
- K. Zhang, from January to December 2022 (O. Balalau, I. Manolescu)

Engineers supervision:

- Simon Ebel (O. Balalau, I. Manolescu)
- Théo Galizzi (O. Balalau, I. Manolescu)
- Tinhinane Medjkoune, at LIG (O. Goga)
- Arnab Singh (Y. Diao)

Intern supervision:

- Antoine Gauquier (M1, IMT Lille), I. Manolescu
- Q. Massonnat (M1, Ecole Polytechnique), O. Balalau and I. Manolescu

10.2.3 Juries

- Oana Balalau: comité de suivi de thèse of Cyril Chhun, Télécom Paris
- Oana Goga: PhD Committee for B. Khalfoun
- Ioana Manolescu: president of the PhD committee of Moussa Kamal Eddine (LIX, Ecole Polytechnique)

10.3 Popularization

10.3.1 Internal or external Inria responsibilities

- Oana Balalau:
  - member of the committee "Moyens calcul Inria"
  - member in the GT comité de centre

10.3.2 Education

O.Balalau, Y.Diao and I. Manolescu reviewed student applications for the Data AI master track at IPP.

10.3.3 Interventions

- Oana Balalau: Participation "Journée après thèse" at Centrale Supelec
- Oana Goga: Participation to podcast La data dans tous ses états
- Ioana Manolescu: Participated in a panel on AI at Centrale Supéléc, organized by a committee of students interested in AI ("Les Automatants")

11 Scientific production

11.1 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] O. Balalau and R. Horincar. 'From the Stage to the Audience: Propaganda on Reddit'. In: EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics. Online, France, 19th Apr. 2021. URL: https://hal.inria.fr/hal-03351621.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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] S. Y. Sahai, O. Balalau and R. Horincar. 'Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions'. In: ACL-IJCNLP 2021 - Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Online, France, 2nd Aug. 2021. URL: https://hal.inria.fr/hal-03351649.

11.2 Publications of the year

International journals

[11] 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'. In: *Information Systems* 104 (1st Oct. 2022), p. 42. DOI: 10.1016/j.is.2021.101846. URL: https://hal.inria.fr/hal-03150441.

[12] S. Minn. 'AI-assisted knowledge assessment techniques for adaptive learning environments'. In: *Computers and Education: Artificial Intelligence* 3 (7th Feb. 2022). DOI: 10.1016/j.caeai.2022.100050. URL: https://hal.inria.fr/hal-03897560.

International peer-reviewed conferences

[13] O. Balalau, S. Ebel, T. Galizzi, I. Manolescu, Q. Massonnat, A. Deiana, E. Gautreau, A. Krempf, T. Pontillon, G. Roux and J. Yakin. 'Fact-checking Multidimensional Statistic Claims in French'. In: TTO 2022 - Truth and Trust Online. Boston [Hybrid Event], United States, 12th Oct. 2022. URL: https://hal.science/hal-03791175.

[14] O. Balalau, S. Ebel, T. Galizzi, I. Manolescu, Q. Massonnat, A. Deiana, E. Gautreau, A. Krempf, T. Pontillon, G. Roux and J. Yakin. 'Statistical Claim Checking: StatCheck in Action (demonstration)'. In: CIKM 2022 - 31st ACM International Conference on Information and Knowledge Management. Atlanta / Hybrid, United States, 17th Oct. 2022. URL: https://hal.inria.fr/hal-03767992.

[15] N. Barret, I. Manolescu and P. Upadhyay. 'Abstra: Toward Generic Abstractions for Data of Any Model'. In: CIKM 2022 - 31st ACM International Conference on Information and Knowledge Management. Atlanta, Georgia / Hybrid, United States, 17th Oct. 2022. URL: https://hal.inria.fr/hal-03767967.

[16] D. Berrebbi, N. Huynh and O. Balalau. 'GraphCite: Citation Intent Classification in Scientific Publications via Graph Embeddings'. In: 2nd International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment. Companion Proceedings of the Web Conference 2022 (WWW '22 Companion). Lyon / Virtual, France, 25th Apr. 2022. URL: https://hal.inria.fr/hal-03648498.

[17] E. Gkiouzepi, A. Andreou, O. Goga and P. Loiseau. 'Collaborative Ad Transparency: Promises and Limitations'. In: 44th IEEE Symposium on Security and Privacy. San Francisco, United States, 2023. URL: https://hal.inria.fr/hal-03916393.

[18] C. Lyu, Q. Fan, F. Song, A. Sinha, Y. Diao, W. Chen, L. Ma, Y. Feng, Y. Li, K. Zeng and J. Zhou. 'Fine-Grained Modeling and Optimization for Intelligent Resource Management in Big Data Processing'. In: VLDB 2022 - 48th International Conference on Very Large Databases. Vol. 15. 11. Sydney, Australia, 5th Sept. 2022. URL: https://hal.inria.fr/hal-03897397.

[19] I. Manolescu. 'Teasing journalistic findings out of heterogeneous sources: a data/AI journey (invited keynote)'. In: *International Conference on Distributed Event-Based Systems (DEBS)*. DEBS '22: The 16th ACM International Conference on Distributed and Event-based Systems. Copenhagen, Denmark: ACM, 22nd Aug. 2022, pp. 1–1. DOI: 10.1145/3524860.3544406. URL: https://hal.inria.fr/hal-03945733.

[20] J.-J. Vie, T. Rigaux and S. Minn. 'Privacy-Preserving Synthetic Educational Data Generation'. In: EC-TEL 2022 - 17th European Conference on Technology Enhanced Learning. Toulouse, France, 12th Sept. 2022. URL: https://hal.inria.fr/hal-03715416.

National peer-reviewed Conferences

[21] A. C. Anadiotis, I. Manolescu and M. Mohanty. 'Integrating Connection Search in Graph Queries'. In: BDA 2022 - 38ème Conférence sur la Gestion de Données - Principes, Technologies et Applications. Clermont-Ferrand, France, 24th Oct. 2022. URL: https://hal.inria.fr/hal-03886320.

[22] M. Ghufran Khan. 'Scalable Analytics on Multi-Streams Dynamic Graphs'. In: BDA conference "Data Management - Principles, Technologies and Applications" 2022. Clermont-Ferrand, France, 24th Oct. 2022. URL: https://hal.inria.fr/hal-03903287.

Conferences without proceedings

[23] A. C. Anadiotis, F. Goasdoué, M. Y. Haddad and I. Manolescu. 'Towards Speeding Up Graph-Relational Queries in RDBMSs'. In: BDA 2022 - 38èmes journées de la conférence BDA « Gestion de Données – Principes, Technologies et Applications. Clermont-Ferrand, France, 24th Oct. 2022. URL: https://hal.inria.fr/hal-03791272.

[24] O. Balalau, S. Ebel, T. Galizzi, I. Manolescu and Q. Massonnat. 'Statistical Claim Checking: StatCheck in Action'. In: 38ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA 2022). Clermont - Ferrand, France, 24th Oct. 2022. URL: https://hal.science/hal-03932371.

[25] N. Barret, I. Manolescu and P. Upadhyay. 'Abstra: Toward Generic Abstractions for Data of Any Model'. In: BDA 2022 - informal publication only. Clermont-Ferrand, France, 24th Oct. 2022. URL: https://hal.inria.fr/hal-03774599.

Scientific book chapters

[26] T. Bouganim, H. Galhardas and I. Manolescu. 'Efficiently Identifying Disguised Missing Values in Heterogeneous, Text-Rich Data'. In: *Transactions on Large-Scale Data- and Knowledge-Centered Systems LI*. Vol. 13410. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 8th Oct. 2022, pp. 97–118. DOI: 10.1007/978-3-662-66111-6_4. URL: https://hal.archives-ouvertes.fr/hal-03817900.

Reports & preprints

[27] A. C. Anadiotis, I. Manolescu and M. Mohanty. *Integrating Connection Search in Graph Queries*. Inria Saclay - Île de France, 4th Jan. 2023. URL: https://hal.inria.fr/hal-03923293.