RESEARCH CENTRE

Inria Saclay Centre at Institut Polytechnique de Paris

IN PARTNERSHIP WITH:

Institut Polytechnique de Paris, CNRS

2023 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

Inria

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-Denisa Balalau [INRIA, ISFP]
- Oana Goga [CNRS, Researcher]

Faculty Member

- Yanlei Diao [Ecole Polytechnique, Professor, HDR]

Post-Doctoral Fellows

- Madhulika Mohanty [INRIA, Post-Doctoral Fellow, until Feb 2023]
- Sein Minn Sein Minn [CNRS, until Jan 2023]

PhD Students

- Nardjes Amieur [CNRS, from Feb 2023]
- Nelly Barret [INRIA]
- Theo Bouganim [INRIA]
- Tom Calamai [Amundi, CIFRE]
- Salim Chouaki [CNRS]
- Asmaa El Fraihi [CNRS, from Feb 2023]
- Qi Fan [Ecole Polytechnique,LIX]
- Mhd Yamen Haddad [INRIA, until Jan 2023]
- Vincent Jacob [Ecole Polytechnique]
- Muhammad Khan [INRIA]
- Kun Zhang [INRIA]

Technical Staff

- Simon Ebel [INRIA, Engineer]
- Theo Galizzi [INRIA, Engineer]
- Madhulika Mohanty [INRIA, Engineer, from Mar 2023]
- Camille Pettineo [INRIA, Engineer, from Feb 2023 until Apr 2023]

Interns and Apprentices

- Abir Benzaamia [CNRS, Intern, from Feb 2023 until Jul 2023]
- Louis Caubet [INRIA, Intern, until Mar 2023]
- Ahmed-Yassine Chraa [ECOLE POLY PALAISEAU, Intern, from Jun 2023 until Jul 2023]
- Nikola Dobricic [INRIA, Intern, from Jul 2023 until Aug 2023]
- Philippe Guyard [Ecole Polytechnique, LIX, Intern, until Mar 2023]
- Amine Lamouchi [Ecole Polytechnique, LIX, Intern, until Mar 2023]
- Melissa Mokhtari [CNRS, Intern, from Feb 2023 until Jul 2023]
- Maria-Natalia Osman-Calescu [INRIA, Intern, from Jul 2023 until Aug 2023]
- Ines Vignal [Ecole Polytechnique, LIX, Intern, until Mar 2023]

Administrative Assistant

- Michael Barbosa [INRIA]

Visiting Scientists

- Andrei-Laurentiu Bornea [Ecole Polytechnique, LIX, from Feb 2023 until Aug 2023]
- Vera Sosnovik [ECOLE POLY PALAISEAU, from Apr 2023 until Jun 2023]

External Collaborators

- Matthieu Beauval [Radio France]
- Mathilde Bouquerel [Radio France, until Nov 2023]
- Estelle Cognacq [Radio France, until Aug 2023]
- Antoine Deiana [Radio France]
- Emilie Gautreau [Radio France]
- Samuel Guimaraes [CNRS, from Nov 2023]
- Stéphane Horel [LE MONDE]
- Antoine Krempf [Radio France, until Mar 2023]
- Chenghao Lyu [Ecole Polytechnique, LIX]
- Isotta Magistrali [IP PARIS, from Oct 2023 until Nov 2023]
- Adrien Maumy [Radio France]
- Gabrielle Mura [IP PARIS, from Oct 2023 until Nov 2023]
- Thomas Pontillon [Radio France, until Apr 2023]
- Gerald Roux [Radio France]
- Saumya Yashmohini Sahai [OSU, until Mar 2023]
- Prajna Devi Upadhyay [BITS PILANI HYDERABAD CAMPUS]
- Joanna Yakin [Radio France, until Apr 2023]

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 natural langauge processing tools and their applications to problems such as argumentation mining, factfulness evaluation and information extraction. We are interested in particular on applications with high social value, such as analysing public discourse with the goal of finding elements that could bias the world view of citizens, such as false claims, fallacious arguments, propaganda, or greenwashing.

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 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.

3.2 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.

3.3 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.

3.4 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.

3.5 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.

3.6 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 SourcesSay AI Chair project, 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

Madhulika Mohanty co-lead the SCOUT action of the Diversity, Equity and Inclusion initiative (website) for the DB research community. This action was in-charge of writing the Code of Ethics and organization guidelines to be used in future DB conferences. This has lead to the publication of [12].

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 8.4.1), and on ad transparency (Section 8.2), goes towards making information sharing on the Web more transparent and more trustworthy.

6 Highlights of the year

6.1 Awards

Oana Goga received the Lovelace-Babbage Award from the French Academy of Science and the French Computer Society. It is given out annually to two researchers below the age of 40 that made significant contributions to Computer Science, from the most theoretical to the most applied.

6.2 Engagement with journalists

In the spring, a journalist and data scientist, Camille Pettineo, worked in the team. Discussing with her, we have devised new, more user-friendly interfaces for our software ConnectionLens, leading to the ConnectionStudio system.

In May 2023, O. Balalau, N. Barret, S. Ebel, T. Galizzi and I. Manolescu presented tools developed within the team (StatCheck, ConnectionStudio, and Abstra) to a meet-up of DataJournos, a data journalism association. Approximately 40 people attended our presentation. This has lead to new collaborations between O. Balalau and ADEME/AEF, on the topic of automatically detecting greenwashing through Natural Language analysis.

In July 2023, N. Barret and M. Mohanty gave a tutorial on ConnectionStudio and how to use it for data journalism, at the "Forum Medias et Développement" organised by CFI, the french media development agency.

In September 2023, O. Balalau, T. Bouganim and I. Manolescu participated to SciCar ("Where Science meets Computer-Assisted Reporting") conference in Dortmund, Germany. We presented our collaboration with S. Horel (Le Monde).

In October 2023, O. Balalau, S. Ebel et T. Galizzi presented StatCheck, our fact-checking tool, at an IA day organized by RadioFrance.

In November 2023, I. Manolescu participated in a debate on "New (AI) tools for media" at the "Médias en Seine" journalism conference.

7 New software, platforms, open data

7.1 New software

7.1.1 ConnectionLens

Name: Integration of heterogeneous data using information extraction

Keyword: Data analysis

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, Madhulika Mohanty

Contact: Manolescu Ioana

7.1.2 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

7.1.3 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 multi-dimensional, 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

7.1.4 ConnectionStudio

Keywords: Heterogeneous Data, Data Exploration

Functional Description: ConnectionStudio integrates highly heterogeneous data into graphs, enriched with extracted entities. Studio users can discover the entities in their data, navigate across connections between datasets, explore and query the data in many ways. The Studio currently supports: CSV, JSON, XML, RDF, text, property graphs, all Office formats, and PDF datasets.

ConnectionStudio is a novel front-end to ConnectionLens, Abstra and PathWays (see also the respective Web sites). Its own novel features are outlined in a CoopIS 2023 article.

URL: https://connectionstudio.inria.fr/

Contact: Ioana Manolescu

7.1.5 FactSpotter

Keywords: Factual Faithfulness, Text generation

Functional Description: We propose a new metric that correctly identifies factual faithfulness, i.e., given a triple (subject, predicate, object), it decides if the triple is present in a generated text. We show that our metric FactSpotter achieves the highest correlation with human annotations on data correctness, data coverage, and relevance. In addition, FactSpotter can be used as a plug-in feature to improve the factual faithfulness of existing models.

Contact: Kun Zhang

Partner: Ecole Polytechnique

7.1.6 PathWays

Keywords: Named entities, Data Journalism

Functional Description: PathWays is a system that is able to find connections between entities, such as people, organizations and locations, across documents. For this task, PathWays starts by loading a set of (potentially heterogeneous) datasets as a data graph (see https://arxiv.org/abs/2012.08830) and then builds a collection graph, i.e. a summary of it (see https://hal.inria.fr/hal-03767967). Based on that collection graph, PathWays enumerates all paths connecting two user-specified entities.

URL: https://team.inria.fr/cedar/projects/pathways/

Contact: Nelly Barret

7.1.7 OpenIEEntity

Name: Open Information Extraction with Entity Focused Constraints

Keyword: Information extraction

Functional Description: This tool takes in input a sentence and outputs the facts contained in the sentence, in the format (subject,predicate,object).

Contact: Oana-Denisa Balalau

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. 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. **ConnectionStudio: a user-friendly data lake for data exploration.** ConnectionStudio [17, 18, 19] is a user-friendly data lake, ingesting structured, semi-structured and un-structured documents (XML, JSON, CSV, RDF, PG, PDFs and Office files). It provides users different ways to explore and query the underlying data. For instance, it provides a set of lake-level statistics about the entities found in the data and Entity-Relationship diagrams showing the datasets structures. Further, it allows users to enumerate and browse a set of (interesting) paths connecting entities of interest in the data. It also provides a querying interface to let users formulate their queries using few variables and inspect the data lake without having to write SQL queries. Finally, users can save, export and share results they produce in the lake, e.g., data journalists can share them in newsrooms.

2. **Integrating Connection Search in Graph Queries.** When graph database users explore unfamiliar graphs, potentially with heterogeneous structure, users may need to find how two or more groups of nodes are connected in a graph, even when users are not able to 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. In this work, we formally showed how to integrate

connecting tree patterns (CTPs, in short) with a graph query language such as GPML, SPARQL or Cypher, leading to Extended Queries (or EQs, in short). 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. To cope with very large search spaces, we propose efficient pruning techniques and formally establish a large set of cases where our best algorithm, MOLESP, is complete even with pruning. Our experiments validate the performance of our algorithms on many synthetic and real-world workloads [15, 16, 23].

3. **Big Graph visualization and exploration.** In this project, we focus on understanding the role played by each node in a graph and building a visualisation tool to explore it node by node, showing the most pertinent next neighbour to visit for a node. This visualisation can be plugged in the the tools we develop in our team. Our challenges are the following. First, nodes in a graph can play several roles, they can be part of an entity, being the root of the entity or describing the entity, such as attributes or values, or they can take part in relations between entities. We have identified and attributed automatically the roles played in a graph by each node. Second, when a user wants to explore the neighbour of a node, we should suggest relevant neighbours. We are working on node scoring functions to determine which neighbours of a node are the most relevant to explore next.

ConnectionLens is available online at: ConnectionLens Gitlab repository, while ConnectionStudio is available at ConnectionStudio Gitlab repository.

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 initial technical novelties of STATCHECK were 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 [2]. In 2023, we have further improved the plaform with: (i) the use of LLM for both the table extraction and the table retrieval; (ii) extending the list of data sources, from which we gather tables, in a extensible and dynamic manner (Focus crawling, incremental crawler) (iii) automatical transcription and analysis of audio files, this new capatibility is perticularly important in the collaboration with Franceinfo, since journalists work a lot with radio feeds; (iv) detection and analysis of propaganda.

8.2 Online targeted advertising

8.2.1 On Detecting Policy-Related Political Ads: An Exploratory Analysis of Meta Ads during the 2022 French Election

Online political advertising has become the cornerstone of political campaigns. The budget spent solely on political advertising in the U.S. has increased by more than 1002017-2018 U.S. election cycle to 1.6 billion during the 2020 U.S. presidential elections. Naturally, the capacity offered by online platforms to micro-target ads with political content has been worrying lawmakers, journalists, and online platforms, especially after the 2016 U.S. presidential election, where Cambridge Analytica has targeted voters with political ads congruent with their personality. To curb such risks, both online platforms and regulators (through the DSA act proposed by the EuropeanCommission) have agreed that researchers, journalists, and civil society need to be able to scrutinize the political ads running on large online platforms. Consequently, online platforms such as Meta and Google have implemented Ad Libraries that contain

information about all political ads running on their platforms. This is the first step on a long path. Due to the volume of available data, it is impossible to go through these ads manually, and we now need automated methods and tools to assist in the scrutiny of political ads. In this work [25], we focus on political ads that are related to policy. Understanding which policies politicians or organizations promote and to whom is essential in determining dishonest representations. This paper proposes automated methods based on pre-trained models to classify ads in 14 main policy groups identified by the Comparative Agenda Project (CAP). We discuss several inherent challenges that arise. Finally, we analyze policy-related ads featured on Meta platforms during the 2022 French presidential elections period.

8.2.2 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 work [22], 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.

8.2.3 Marketing to Children Through Online Targeted Advertising: Targeting Mechanisms and Legal Aspects

Many researchers and organizations, such as WHO and UNICEF, have raised awareness of the dangers of advertisements targeted at children. While most existing laws only regulate ads on televi- sion that may reach children, lawmakers have been working on extending regulations to online advertising and, for example, forbid (e.g., the DSA) or restrict (e.g., the COPPA) advertising based on profiling to children. At first sight, ad platforms such as Google seem to protect chil- dren by not allowing advertisers to target their ads to users that are less than 18 years old. However, this work [24] shows that other targeting features can be exploited to reach children. For example, on YouTube, advertisers can target their ads to users watching a particular video through placement-based targeting, a form of con- textual targeting. Hence, advertisers can target children by simply placing their ads in children-focused videos. Through a series of ad experiments, we show that placement-based targeting is possible on children-focused videos and, hence, enables marketing to chil- dren. In addition, our ad experiments show that advertisers can use targeting based on profiling (e.g., interest, location, behavior) in combination with placement-based advertising on children-focused videos. We discuss the lawfulness of these two practices with re- spect to DSA and COPPA. Finally, we investigate to which extent real-world advertisers are employing placement-based targeting to reach children with ads on YouTube. We propose a measurement methodology consisting of building a Chrome extension able to capture ads and instrumenting six browser profiles to watch children-focused videos. Our results show that 7we test use placement-based targeting. Hence, targeting children with ads on YouTube is not only hypothetically possible but also occurs in practice. We believe that the current legal and technical solutions are not enough to protect children from harm due to online advertising. A straightforward solution would be to forbid placement-based advertising on children-focused content.

8.2.4 Understanding the privacy risks of popular search engine advertising systems

Privacy-focused search engines such as DuckDuckGo, StartPage, and Qwant promote a strategy of respecting users' privacy and promise not to track users' search and browsing behavior. However, they rely

on advertising for revenue, using Microsoft's (DuckDuckGo and Qwant) or Google's (StartPage) advertising systems. Moreover, these search engines are often silent or ambiguous on the privacy properties of the ads that appear on their search page. Our research [21] delves into the privacy properties of advertising systems used by these search engines. Our findings reveal that the privacy protections of private search engines do not sufficiently cover their advertising systems. Although these search engines refrain from identifying and tracking users and their ad clicks, the presence of ads from Google or Microsoft subjects users to the privacy-invasive practices performed by these two advertising platforms. When users click on ads on private search engines, they are often identified and tracked either by Google, Microsoft, or other third parties, through bounce tracking and UID smuggling techniques.

8.3 How effective are tracking restrictions: A case study on Meta

Growing concern over digital privacy has led to the widespread use of tracking restriction tools such as ad blockers, Virtual Private Networks (VPN), and privacy-focused web browsers. Despite these efforts, advertising companies continuously innovate to overcome these restrictions. Recently, advertising platforms like Meta have been promoting server-side tracking solutions to bypass traditional browser-based tracking restrictions.

We explore how server-side tracking technologies can link website visitors with their user accounts on Meta products. The goal is to assess the effectiveness and accuracy of employing this technology, as well as the effect of tracking restrictions on online tracking. Our methodology involves a series of experiments where we integrate Meta's client-side tracker (the Meta Pixel) and server-side technology (the Conversions API) on different web pages. We then drive traffic to these pages and evaluate the success rate of linking website visitors to their profiles on Meta products.

8.4 Improving the quality of public debate with AI

8.4.1 Argumentation mining and applications to public discourse

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 [3]. In the past, we have worked on propaganda [3] and fallacy detection [11]. We continue this work with a CIFRE PhD, 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. We are currently collaborating with ADEME and AEF Info on the topic of greenwashing detection.

8.4.2 Finding conflicts of interest via 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 often named entities, we explore how to inject constraints in the extraction through constrained inference and constraint-aware training. An important use case that we want to pursue next is automatically creating a knowledge base of relations between scientists and companies, i.e. identifying conflict-of-interest between the scientists and funding bodies, where the named entities are the names of scientists and companies, and the relation describes

the conflict of interest between them. Our work [26] 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

8.4.3 Detecting hallucinations in Large Language Models

This research [27] investigates the extent to which the Graph-to-Text (G2T) generation problem is addressed in existing datasets and the performance of metrics in text comparison. A key focus is on a new metric, FactSpotter, developed to assess factual faithfulness in generated texts from G2T models. FactSpotter is shown to correlate highly with human annotations in terms of data correctness, coverage, and relevance. It functions as a plugin feature to enhance the factual accuracy of existing models and examines the current challenges in G2T datasets.

FactSpotter, initially developed for evaluating the factual faithfulness of Graph-to-Text generation, has its potential for expansion into a more general text similarity metric. This would enable it to assess a wider range of text generation tasks, extending its applicability to various narrative forms and data-driven journalism.

8.5 Efficient Big Data analytics

8.5.1 Efficient and robust active learning methods for interactive database exploration

There is an increasing gap between 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 the user retrieve high-value content from data more effectively. In this work [14], we propose an interactive data exploration system as a new database service, using an approach called "explore-by-example." Our new system is designed to assist the user in performing highly effective data exploration while reducing the human effort in the process. We cast the explore-by-example problem in a principled "active learning" framework. However, traditional active learning suffers from two fundamental limitations: slow convergence and lack of robustness under label noise. To overcome the slow convergence and label noise problems, we bring the properties of important classes of database queries to bear on the design of new algorithms and optimizations for active learning-based database exploration. Evaluation results using real-world datasets and user interest patterns show that our new system, both in the noise-free case and in the label noise case, significantly outperforms state-of-the-art active learning techniques and data exploration systems in accuracy while achieving the desired efficiency for interactive data exploration.

8.5.2 Multi-Objective Optimization for Spark-based Data Analytics

Spark has been widely used for data analytics in the cloud. Determining an optimal configuration of a Spark physical plan based on user-specified objectives is a complex task. It is challenging from three aspects. Firstly, a Spark physical plan, or query, can be represented as a Directed Acyclic Graph (DAG) of “query stages,” where parameters of each stage are controlled under the granularity of the query (i.e. Spark-context parameters, e.g. resources are shared among all stages) and the granularity of the stage (i.e. different among different stages). The correlation of parameters under multiple granularity makes the performance tuning of a query more complicated. Secondly, the parameters of each stage face timing constraints. Spark-context parameters should be set at compile time and cannot change during runtime, while stage-level parameters can be modified during runtime. Thirdly, Multi-Objective Optimization (MOO) is necessary when there are multiple potentially conflicting, user performance objectives such as latency and cost. This work focuses on the algorithm design to return Pareto optimal configurations for a query with parameters under multi-granularity control and different timing constraints. It captures tradeoffs among various objectives and recommends an optimal configuration based on user preferences. The expectation is to provide recommendations for all stages in a query within a few seconds.

8.5.3 Scalable Analytics on Multi-Streams Dynamic Graphs

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. In this project [28], we 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.

8.6 Explainable Anomaly Detection on Multivariate Time Series in the AIOps Domain

The widespread adoption of Internet-based services by software companies, as well as the scale and complexity at which they operate, have made incidents in their IT operations increasingly more likely, diverse and impactful. This has led to the rapid development of a central aspect of the "Artificial Intelligence for IT Operations" (AIOps) domain, focusing on detecting abnormal patterns in vast amounts of multivariate time series (MTS) data generated by service entities. Although numerous MTS anomaly detection methods have been developed, the state of the art still presents some limitations due to the unique challenges posed by AIOps. These challenges include 1) the presence of complex, noisy and diverse normal behaviors, 2) the wide variety of anomaly types and difficulty in providing detailed anomaly labels, and 3) the need to generalize to a wide variety of normal behaviors for the monitored entities.

Our research focused on designing new anomaly detection methods to address these AIOps challenges, mainly through explicit context generalization and weak supervision, as well as conducting thorough experiments to demonstrate their superiority. Concurrently, we continued developing our data science pipeline for explainable anomaly detection over MTS, to further facilitate the design and benchmarking of new techniques for the community.

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 €)

Ioana Manolescu is the local PI for the Inria partner in the project "ELIAS - European Lighthouse of AI for Sustainability" (2,800,000€). Oana Goga is also involved.

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 €)
- Ioana Manolescu is the local PI for Inria Saclay in CQFD (2019-2024), an ANR project coordinated by F. Ulliana (U. Montpellier). Its research aims at investigating efficient data management methods for ontology-based access to heterogeneous databases (polystores).

- Ioana Manolescu is the PI of SourcesSay (2020-2024), 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

- ANRT Project: CIFRE Amundi, advised by Oana 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

Ioana Manolescu is the PI for the region-funded project DIM AI4IDF. 2023 has been its third and last year; in particular, it has supported the PhD of Nelly Barret.

Ioana Manolescu, Oana Balalau have an Action Exploratoire Inria project (2023-2025), "JoDaIA: Harnessing Data and AI for Journalism".

Hi!Paris Collaborative Project (2022-2024) coordinated by Oana Balalau and J.Romero (Télécom SudParis).

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair Oana Balalau and Ioana Manolescu have been involved in organizing the Hi!Paris summer school in AI for Science and Society, in July 2023, at HEC, Jouy en Josas. Ioana Manolescu has co-chaired the school, and Oana Balalau has been a session chair.

Member of the organizing committees

- Ioana Manolescu has co-organized the VLDB Summer School in Data Management in Cluj, Romania, in July 2023.
- Oana Goga and Ioana Manolescu were co-organizers of the first interdisciplinary workshop "Infox sur Seine", bringing social scientists, political scientists and computer scientists together around topics of dis- and mis-information, March 22-23, 2023.
- Yanlei Diao was the organizer of the ERC BigFastData workshop, a two day event at Ecole Polytechnique, October 5-6, 2023, with invited speakers all over Europe and USA.

10.1.2 Scientific events: selection

Chair of conference program committees Ioana Manolescu: Tutorial chair at the EDBT 2023 Conference, Meta-reviewer (Industry and Applications track) at the IEEE ICDE 2023 conference

Member of the conference program committees

- Oana Balalau: ACL 2023, EMNLP 2023, CSCW 2024, BDA2023
- Oana Goga: The Web Conference (2023), USENIX Security (2023), ConPro (2023), FAccT (2023), IC2S2 (2023)

- Ioana Manolescu: ACM SIGMOD 2023, BDA (Demonstrations) 2023
- Madhulika Mohanty: CoDS-COMAD 2023, WSDM 2023, SIGMOD Availability and Reproducibility 2023, DSAA 2023, SIGIR (Demo) 2023, ICWE 2023, IJCAI 2023, AIMLSystems 2023.
- Kun Zhang: ICANN 2023

10.1.3 Journal

Member of the editorial boards Ioana Manolescu joined the editorial board of the VLDB Journal in September 2023

Reviewer - reviewing activities Madhulika Mohanty: IEEE Transactions on Knowledge and Data Engineering (TKDE), Knowledge and Information Systems (KAIS)

10.1.4 Invited talks

- Ioana Manolescu:
  - Keynote “All your digital data sources in one place: the SourcesSay ANR/DGA project” at the Infox sur Seine workshop, on March 22, 2023
  - Keynote at the yearly meeting of the "RADIA (Raisonnement, Données et IA)" CNRS research group, in Strasbourg, on July 2, 2023.
- Oana Balalau, Théo Bouganim and Ioana Manolescu were invited to present, together with Stéphane Horel (Le Monde) and Gary Fooks and Tom Mills (UK), our joint work on use ConnectionLens to build a database about conflicts of interest in the biomedical domain, at SciCAR 2023 (Sept 29-30, 2023, in Dortmund, Germany).
- Oana Balalau presented her work at AIvolution, a workshop on AI and digital technologies at the European Parliment in Brussels.
- Oana Goga:
  - (N) Musée des Arts et Métiers – L’aventure des inventions, Feb 2024, general audience, invited along very distinguished speakers (I) WSTEAM Panel AICCSA, Dec 2023 – non research
  - (I) Plenary talk for AIvolution, European Parliament, Bruxelles, November 2023
  - (I) Plenary talk at the OECD event Tackling disinformation: Strengthening democracy through information integrity, OECD, Paris, November 2023, 400+ participants, I presented along government officials with high functions such as chancellors and state secretaries
  - (I) Plenary talk at Atelier 2 - Projet de recherche collective - Vers un droit neuroéthique?, Paris, November 2023
  - (I) Keynote at Journées Nationales GDR Securité, Paris, June 2023
  - (F) Plenary talk at VIGINUM on “Rencontres et débats autour des manipulations de l’information”, Paris, June 2023
  - (F) Plenary talk at Conférence Arcom x CNNum on “Role de la société civile dans la régulation numérique nationale et européenne”, Paris, June 2023
  - (I) Keynote at 7th ASF/RSD Winter school, Sept Laux, January 2023
  - (N) Seminar DataScale master, Versailles, Dec 2023
  - (N) LIX Lab Seminar, Palaiseau, April 2023
  - (N) Inria PETRUS Team Seminar, Versailles, April 2023
  - (N) PEREN–Pole d’Expertise de la Regulation Numerique (Academic Seminar), Quai de Bercy, Paris, March 2023 (N) Université Dauphine (Seminar “Responsabilité Sociale des Algorithmes”), Paris, February 2023

- Yanlei Diao:
  - SWIFT Global AI Forum, February 8th, 2023
  - Database Summit, MIT, October 20, 2023

10.1.5 Leadership within the scientific community

- Ioana Manolescu became the president of the steering committee of the BDA association "Communauté Francophone en Gestion de données : Principes, Technologies et Applications" in November 2023.
- Yanlei Diao served as a member of ACM SIGMOD Award Committee

10.1.6 Scientific expertise

- Oana Goga is working as an External Expert with contract with the European Commission (DG Connect and ECAT), to advise on the Delegated Act addressing the implementation of the DSA Article 40 that gives vetted researchers access to data from very large online platforms and very large search engines.
- Oana Goga became a member of Conseil Scientifique of Regalia (2023)
- Oana Goga became a 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 (2022-2023).
- June 2022, Oana Goga was part of a team that answered the ARCOM “Public consultation on access to data from online platforms”. The new Digital Service Act provides legal grounds for researchers to ask for data from online platforms. The ARCOM proposed several procedures to access this data. We wrote a document highlighting the problems with the current procedures. The fact that together with Beatrice Roussillon and Juliette Senechal we organized a community of social scientists, computer scientists, and legal scholars around online platforms allowed us to be in the perfect position to provide clear and informed answers to the consultation.

10.1.7 Research administration

- Ioana Manolescuis a member of the Comité de Direction of LIX, the Computer Science lab of Ecole Polytechnique to which the team belongs.
- Oana Goga is a co-leader of the Action PLATFORM in GDR CIS.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

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

- INF473G “Mining, learning and reasoning on Web Graphs”, L3, Ecole Polytechnique
- INF583 “Systems for Big Data”, M1, Ecole Polytechnique

Nelly Barret: CSE204 (Machine Learning), Bachelor 2nd year, Ecole Polytechnique, 20h TP.
Salim Chouaki:

- Labs - Programming project in Python - ENSAE
- Labs - Programming in C++ - ENSAE

Yanlei Diao: Instructor, “Systems for Big Data”, M1, Ecole Polytechnique
Ioana Manolescu

- Part-time professor (50%) at Ecole Polytechnique. Courses, labs and TDs in INF553 (Database Management Systems); in charge of the M1 Internship program in Artificial Intelligence and Data Science (INF592); in charge of the Artificial Intelligence and Data Science program (M1) at Ecole Polytechnique
- Taught TPT DATAIA921 (Architectures for Big Data), Institut Mines Télécom, 12h of lecture.

Madhulika Mohanty:

- TPT DATAIA921 (Architectures for Big Data), M2, Institut Mines Télécom, 9h TP.
- INF540 (Databases), M1MATHAPP, Ecole Polytechnique, 35h EQTD.

10.2.2 Supervision

PhD supervision: The team has supervised the following PhDs students:

- Nardjes AMIEUR, from February 2023 to December 2023 (Oana Goga)
- Nelly Barret, from January to December 2023 (Ioana Manolescu)
- Théo Bouganim, January to December 2023 (Ioana Manolescu, E. Pietriga (Inria ILDA))
- Tom Calamai, from January to December 2023 (Oana Balalau, F.Suchanek)
- Salim Chouaki, from January to December 2023 (Oana Goga)
- Asmaa EL FRAIHI, from February 2023 to December 2023 (Oana Goga)
- Qi Fan, from January to December 2023 (Yanlei Diao)
- Antoine Gauquier, from October to December 2023, at ENS Paris (Ioana Manolescu, Pierre Senellart)
- Ghufran Khan, from January to December 2023 (A. Anadiotis, Ioana Manolescu)
- Vincent Jacob, from January to December 2023 (Yanlei Diao)
- Chenghao Lyu, from January to December 2023 (Yanlei Diao, visiting PhD From UMass Amherst)
- Vera Sosnovik, at LIG, from January to July 2023 (Oana Goga)
- Kun Zhang, from September to December 2023 (Oana Balalau, Ioana Manolescu)

Engineers supervision:

- Tinhinane Medjkoune, at LIG, Jan. 2022 – Jan. 2023 (Oana Goga)
- Théo Galizzi, Simon Ebel (Ioana Manolescu, Oana Balalau)
- Arnab Singh (Yanlei Diao)

Intern supervision: The team has supervised the following interns:

- Abir Benzaamia (M2, ESI Alger) "Privacy-preserving recommender systems on mobile phones", Oana Goga
- Nikola Dobricic (Bachelor 3A, Ecole Polytechnique) "User-friendly and expressive interfaces for querying data in ConnectionStudio", Ioana Manolescu, Simon Ebel
- Melissa Mokhtari (M2, ESI Alger) "Studying how (mis)information spreads in Facebook groups", Oana Goga
- Maria Osman-Calescu (Bachelor 3A, Ecole Polytechnique): "Deduplicating entities based on data cleaning functions",Ioana Manolescu

- Tu Nguyen (Bachelor 2A, Ecole Polytechnique): "Open information extraction with applicatins to conflict of interest extraction", Oana Balalau, Ioana Manolescu, P. Upadhyay

Part-time project supervision The team has supervised the following part-time research projects:

- Junyuan Wang, Zhuoya Zang (L3, Ecole Polytechnique), "Mapping and analysing statistical sites", Simon Ebel and Ioana Manolescu

- Isotta Magistrali, Gabriele Mura (L3, Ecole Polytechnique), "Analysing AI-automated fact-checks", Théo Galizzi, Ioana Manolescu,Oana Balalau

- Tudor Enache (L3, Ecole Polytechnique), "From Simple Graphs to Property Graphs", Nelly Barret, Madhulika Mohanty, Ioana Manolescu

- Shay Pripstein (L3, Ecole Polytechnique), "Targeted Data Acquisition from Open Data Repositories", Madhulika Mohanty, Nelly Barret,Ioana Manolescu

- Andrei Bornea (L2, Ecole Polytechnique), "Open information extraction with constraints", Oana Balalau, Julien Romero

10.2.3 Juries

- Ioana Manolescu has participated to a hiring committee for a Database/AI professor at Université de Grenoble.

- Ioana Manolescu has been part of the PhD defense committee of Wafaa Al-Husaini, at Univ. Rennes 1. She has been a member of a comité de suivi of François Maine, at U. Paris Sorbonne.

- Oana Balalau and Yanlei Diao have been part of the PhD defense committee of Katia Antonenko, Ecole Polytechnique. Yanlei Diao was the president of the jury.

- Oana Balalau has been a jury in the comité de suivi for: Rajaa El Hamdani (Télécom Paris), Cyril Chhun (Télécom Paris), and Jonathan Colin (Université de Saclay)

- Oana Goga was in the following juries:

  - Institute Evaluation Committee: Evaluation of the Max Planck Graduate Center Computer Science (Jan 2023)
  - Award Committee: CNIL-Inria Privacy Protection Award (2023),
  - Ph.D. Committee: Moitree Basu (Jan 2024), Anne Josiane (May 2023) – as jury member
  - Grants Commitees: ANR - Comité d'évaluation CE39 (2023), Max Planck Society Project (2023), German National Research Center for Applied Cybersecurity ATHEN (2023)

10.3 Popularization

10.3.1 Internal or external Inria responsibilities

- Ioana Manolescu has been elected a member at the Inria Commission d'Evaluation in July 2023. This newly elected commission started work in September 2023.

- Oana Balalau was a member of the GT comité de centre, Inria Saclay's Scientific Commission, and the comittee Moyens calcul Inria.

10.3.2 Articles and contents

- An interview with Ioana Manolescu on fact-checking and journalistic fact-checking appeared in "20 minutes": Le monde ne se privera jamais d'un journaliste qui s'y connaît dans un domaine, pour Ioana Manolescu
- StatCheck, the statistic fact-checking tool we develop and which RadioFrance uses, has been featured in "La revue des médias", in an article by Xavier Eutrope: Intelligence artificielle et médias : cinq utilisations, au-delà de ChatGPT
- Ioana Manolescu participated to a debate on "New (AI) tools for media" at the "Médias en Seine" journalism conference in November 2023: Table ronde De nouveaux outils IA et data au service de l'info
- Ioana Manolescu participated to the debate "IA, fake news, and young people's distrust of the media" on FranceInfo and Twitch, on Nov 22, 2023. Replay: Intelligence artificielle, défiance des jeunes, fake news : les médias à l'heure des grandes crises
- Simon Ebel and Théo Galizzi participated to Capcom 2023, presenting StatCheck, developed in collaboration with RadioFrance
- Oana Goga did an interview for Le Monde: Vidéos pour enfants : un ciblage publicitaire qui contourne la réglementation

10.3.3 Interventions

-
- Nelly Barret: "User-oriented exploration of semi-structured datasets" @ LISN (Univ. Paris Saclay) in Oct 2023
- Madhulika Mohanty: "Effective Exploration of Graph-structured Data" at BD, LIP6 in Apr 2023; BD, LIRIS in Mar 2023 and LaHDaK, LISN in May 2023.
- Oana Balalau: "NLP for Journalism: Current Progress and Open Challenges", at Centrum Wiskunde Informatica (CWI) in September 2023
- Oana Balalau: "CS for social good: 2018 vs 2023", at Max Planck Institute for Informatics, in September 2023

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, 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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

[12] S. Amer-Yahia, D. Agrawal, Y. Amsterdamer, S. Bhowmick, A. Bonifati, R. Borovica-Gajic, J. Camacho-Rodríguez, B. Catania, P. Chrysanthis, C. Curino, J. Darmont, G. Dobbie, A. El Abbadi, A. Floratou, J. Freire, A. Jindal, V. Kalogeraki, S. Maiyya, A. Meliou, M. Mohanty, B. Omidvar-Tehrani, F. Özcan, L. Peterfreund, W. Rahayu, S. Sadiq, S. Sellami, U. Sirin, W.-C. Tan, B. Thuraisingham, Y. Tian, P. Tözün, G. Vargas-Solar, N. Yadwadkar, V. Zakhary and M. Zhang. 'Diversity, Equity and Inclusion Activities in Database Conferences: A 2022 Report'. In: *SIGMOD record* 52.2 (10th Aug. 2023), pp. 38–42. DOI: 10.1145/3615952.3615964. URL: https://hal.science/hal-04271933.

[13] L. Di Palma, D. Yanlei and L. Anna. 'Efficient Version Space Algorithms for Human-in-the-loop Model Development'. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 18.3 (12th Jan. 2024), pp. 1–49. DOI: 10.1145/3637443. URL: https://inria.hal.science/hal-04414855.

[14] E. Huang, Y. Diao, A. Liu, L. Peng and L. D. Palma. 'Efficient and robust active learning methods for interactive database exploration'. In: *The VLDB Journal* (16th Nov. 2023). DOI: 10.1007/s00778-023-00816-x. URL: https://inria.hal.science/hal-04414815.

International peer-reviewed conferences

[15] A. C. Anadiotis, I. Manolescu and M. Mohanty. 'Integrating Connection Search in Graph Queries'. In: ICDE 2023 - 39th IEEE International Conference on Data Engineering. Anaheim (CA), United States, 3rd Apr. 2023. URL: https://inria.hal.science/hal-04110779.

[16] A. C. Anadiotis, I. Manolescu and M. Mohanty. 'More power to SPARQL: From paths to trees'. In: *Lecture Notes in Computer Science*. ESWC 2023 - Extended Semantic Web Conference. Vol. LNCS-13870. The Semantic Web - 20th International Conference, ESWC 2023. Hersonissou, Crete, Greece, 28th May 2023. URL: https://inria.hal.science/hal-04102807.

[17] N. Barret, S. Ebel, T. Galizzi, I. Manolescu and M. Mohanty. 'User-friendly exploration of highly heterogeneous data lakes'. In: CoopIS 2023 - International Conference on Cooperative Information Systems. Groningen, Netherlands, 30th Oct. 2023. URL: https://hal.science/hal-04185938.

[18] N. Barret, A. Gauquier, J.-J. Law and I. Manolescu. 'Exploring heterogeneous data graphs through their entity paths'. In: ADBIS 2023 - 27th European Conference on Advances in Databases and Information Systems. Barcelona, Spain, 4th Sept. 2023. URL: https://hal.science/hal-04131977.

[19] N. Barret, A. Gauquier, J.-J. Law and I. Manolescu. 'PathWays: entity-focused exploration of heterogeneous data graphs'. In: ESWC 2023 - 20th European Semantic Web Conference. Hersonissos (Crete), Greece, 28th May 2023. URL: https://hal.science/hal-04103293.

[20] N. Barret, I. Manolescu and P. Upadhyay. 'Computing Generic Abstractions from Application Datasets'. In: *OpenProceedings*. EDBT 2024 - 27th International Conference on Extending Database Technology. Vol. 27. Paestum, Italy, 25th Mar. 2024, pp. 94–107. URL: https://hal.science/hal-04131974.

[21] S. Chouaki, O. Goga, H. Haddadi and P. Snyder. 'Understanding the Privacy Risks of Popular Search Engine Advertising Systems'. In: ACM IMC 2023 - Internet Measurement Conference 2023. Montréal, Canada, 24th Oct. 2023. DOI: 10.1145/3618257.3624823. URL: https://inria.hal.science/hal-04228304.

[22] E. Gkiouzepi, A. Andreou, O. Goga and P. Loiseau. 'Collaborative Ad Transparency: Promises and Limitations'. In: SP 2023 - 44th IEEE Symposium on Security and Privacy. San Francisco, United States, 22nd May 2023. URL: https://inria.hal.science/hal-03916393.

[23] I. Manolescu and M. Mohanty. 'Full-Power Graph Querying: State of the Art and Challenges'. In: VLDB 2023 - 49th International Conference on Very Large Data Bases. Vancouver, Canada, 28th Aug. 2023. DOI: 10.14778/3611540.3611577. URL: https://inria.hal.science/hal-04199455.

[24] T. Medjkoune, O. Goga and J. Sénéchal. 'Marketing to Children Through Online Targeted Advertising: Targeting Mechanisms and Legal Aspects'. In: The 2023 ACM SIGSAC Conference on Computer and Communications Security. Copenhague, Denmark: ACM, 1st Nov. 2023. DOI: 10.1145/3576915.3623172. URL: https://hal.univ-lille.fr/hal-04311598.

[25] V. Sosnovik, R. Kessi, M. Coavoux and O. Goga. 'On Detecting Policy-Related Political Ads: An Exploratory Analysis of Meta Ads in 2022 French Election'. In: WWW 2023 - The ACM Web Conference 2023. Proceedings of the ACM Web Conference 2023. Austin Texas, United States: ACM, 30th Apr. 2023, pp. 4104–4114. DOI: 10.1145/3543507.3583875. URL: https://hal.science/hal-04129915.

[26] P. Upadhyay, O. Balalau and I. Manolescu. 'Open Information Extraction with Entity Focused Constraints'. In: EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia, 2nd May 2023. URL: https://inria.hal.science/hal-03980046.

[27] K. Zhang, O. Balalau and I. Manolescu. 'FactSpotter: Evaluating the Factual Faithfulness of Graph-to-Text Generation'. In: Findings of EMNLP 2023 - Conference on Empirical Methods in Natural Language Processing. Singapore, Singapore, 6th Dec. 2023. URL: https://hal.science/hal-04257838.

National peer-reviewed Conferences

[28] A. Anadiotis, M. Ghufran Khan and I. Manolescu. 'Scalable Analytics on Multi-Streams Dynamic Graphs'. In: BDA 2023 - 39th Conference on Data Management – Principles, Technologies and Applications. Montpellier, France, 5th July 2023. DOI: 10.1145/nnnnnnn.nnnnnnn. URL: https://hal.science/hal-04212814.

Reports & preprints

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

Other scientific publications

[30] I. Manolescu. *Understanding and Querying Data Regardless of the Data Model*. Cluj-Napoca, Romania, 26th July 2023. URL: https://inria.hal.science/hal-04190926.