RESEARCH CENTRE

Grenoble - Rhône-Alpes

IN PARTNERSHIP WITH:

Institut polytechnique de Grenoble, CNRS, Université de Grenoble Alpes

2021 ACTIVITY REPORT

Project-Team TYREX

Types and Reasoning for the Web

IN COLLABORATION WITH: Laboratoire d'Informatique de Grenoble (LIG)

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Contents

Project-Team TYREX

Creation of the Project-Team: 2014 July 01

Keywords

Computer sciences and digital sciences

A2.1.1. – Semantics of programming languages

A2.1.4. – Functional programming

A2.1.7. – Distributed programming

A2.1.10. – Domain-specific languages

A2.2.1. – Static analysis

A2.2.4. – Parallel architectures

A2.2.8. – Code generation

A2.4. – Formal method for verification, reliability, certification

A3.1. – Data

A3.1.1. – Modeling, representation

A3.1.2. – Data management, quering and storage

A3.1.3. – Distributed data

A3.1.6. – Query optimization

A3.1.9. – Database

A3.1.10. – Heterogeneous data

A3.1.11. – Structured data

A3.2.1. – Knowledge bases

A3.2.2. – Knowledge extraction, cleaning

A3.2.6. – Linked data

A3.3.3. – Big data analysis

A3.4. – Machine learning and statistics

A3.4.1. – Supervised learning

A6.3.3. – Data processing

A7. – Theory of computation

A7.1. – Algorithms

A7.2. – Logic in Computer Science

A9.1. – Knowledge

A9.2. – Machine learning

A9.7. – AI algorithmics

A9.8. – Reasoning

A9.10. – Hybrid approaches for AI

Other research topics and application domains

B2. – Health

B6.1. – Software industry

B6.5. – Information systems

B9.5.1. – Computer science

B9.5.6. – Data science

B9.7.2. – Open data

B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientists

- Pierre Genevès [Team leader, CNRS, Senior Researcher, HDR]
- Nabil Layaïda [Inria, Senior Researcher, HDR]

Faculty Members

- Angela Bonifati [Univ Claude Bernard, Professor, until Feb 2021, HDR]
- Ugo Comignani [Institut polytechnique de Grenoble, Associate Professor]
- Nils Gesbert [Institut polytechnique de Grenoble, Associate Professor]

PhD Students

- Sarah Chlyah [Inria]
- Amela Fejza [Univ Grenoble Alpes, until Sep 2021]
- Muideen Lawal [Univ Grenoble Alpes, until Jul 2021]
- Luisa Werner [Univ Grenoble Alpes]

Technical Staff

- Thomas Calmant [Inria, Engineer, until Nov 2021]

Interns and Apprentices

- Hadi Dayekh [Inria, from Feb 2021 until Aug 2021]

Administrative Assistant

- Helen Pouchot-Rouge-Blanc [Inria]

External Collaborators

- Laurent Carcone [ERCIM]
- Amela Fejza [Univ Grenoble Alpes, from Oct 2021]

2 Overall objectives

2.1 Objectives

We work on the foundations of the next generation of data analytics and data-centric programming systems. These systems extend ideas from programming languages, artificial intelligence, data management systems, and theory. Data-intensive applications are increasingly more demanding in sophisticated algorithms to represent, store, query, process, analyse and interpret data. We build and study data-centric programming methods and systems at the core of artificial intelligence applications. Challenges include the robust and efficient processing of large amounts of structured, heterogeneous, and distributed data.

On the data-intensive application side, our current focus is on building efficient and scalable analytics systems. Our technical contributions particularly focus on the optimization, compilation, and synthesis of information extraction and analytics code, in particular with large amounts of data.

On the theoretical side, we develop the foundations of data-centric systems and analytics engines with a particular focus on the analysis and typing of data manipulations. We focus in particular on the foundations of programming with distributed data collections. We also study the algebraic and logical foundations of query languages, for their analysis and their evaluation.

3 Research program

3.1 Foundations for Data Manipulation Analysis: Logics and Type Systems

We develop methods for the static analysis of queries and programs that manipulate structured data (such as trees or graphs). One originality of our research is that we develop type-systems based on decision procedures for expressive logics. One major scientific difficulty here consists in dealing with problems of high computational complexity (sometimes even close to the frontier of decidability), and therefore in finding useful trade-offs between programming expressivity, complexity, succinctness, algorithmic techniques and effective implementations.

3.2 Algebraic Foundations for Optimization of Information Extraction

We explore and develop intermediate languages based on algebraic foundations for the representation, characterization, transformations and compilation of queries. In particular, we investigate two lines of algebraic foundations. First, we study extensions of the relational algebra for optimizing expressive recursive queries. Second, we also explore monad comprehensions and in particular monoid calculi for the generation of efficient and scalable code on big data frameworks. When transforming and optimizing algebraic terms, we rely on cost-based searches of equivalent terms. We thus develop cost models whose purpose is to estimate the time, space and network costs of query evaluation. One difficulty is to estimate these costs in architectures where data and computations are distributed, and where the modeling of data transfers is essential.

4 Application domains

4.1 Querying Large Graphs

Increasingly large amounts of graph-structured data become available. The methods we develop apply for the efficient evaluation of graph queries over large — and potentially distributed — graphs. In particular, we consider the SPARQL query language, which is the standard language for querying graphs structured in the Resource Description Format (RDF). We also consider other increasingly popular graph query languages such as Cypher queries for extracting information from property graphs.

We compile graph queries into lower-level distributed primitives found in big data frameworks such as Apache Spark, Flink, etc. Applications of graph querying are ubiquitous and include: large knowledge bases, social networks, road networks, trust networks and fraud detection for cryptocurrencies, publications graphs, web graphs, recommenders, etc.

4.2 Predictive Analytics for Healthcare

One major expectation of data science in healthcare is the ability to leverage on digitized health information and computer systems to better apprehend and improve care. The availability of large amounts of clinical data and in particular electronic health records opens the way to the development of quantitative models for patients that can be used to predict health status, as well as to help prevent disease and adverse effects.

In collaboration with the CHU Grenoble, we explore solutions to the problem of predicting important clinical outcomes such as patient mortality, based on clinical data. This raises many challenges including dealing with a very high number of potential predictor variables and resource-consuming data preparation stages.

5 New software and platforms

We have continued the development of the main research prototypes MuIR and MedAnalytics and extended them with new features. In particular, we have made progress on the core query optimizer, and on the implementation of Dist-μ-RA.

Dist-μ-RA [10],is a distributed evaluator of recursive graph queries such as UCRPQs. Dist-μ-RA builds on the recursive relational algebra and extends it with evaluation plans suited for the distributed setting. The tool supports an expressive language capturing high-level graph queries while providing efficiency at scale and reducing communication costs on platforms such as Spark with HDFS or Postgres data backends. It automatically generates appropriate distributed optimal execution plans. Experimental results on both real and synthetic graphs show that the tool is effective on very large graphs compared to existing systems such as BigDataLog or Myria.

5.1 New software

5.1.1 MedAnalytics

Keywords: Big data, Predictive analytics, Distributed systems

Functional Description: We implemented a method for the automatic detection of at-risk profiles based on a fine-grained analysis of prescription data at the time of admission. The system relies on an optimized distributed architecture adapted for processing very large volumes of medical records and clinical data. We conducted practical experiments with real data of millions of patients and hundreds of hospitals. We demonstrated how the various perspectives of big data improve the detection of at-risk patients, making it possible to construct predictive models that benefit from volume and variety. Parts of this prototype implementation are described in the publications DSAA'18, Big Data'18, CHIL'21, UAI'21.

Publications: hal-01517087, hal-01877742, hal-03124966, hal-03125018, hal-03160473, hal-03066941, hal-03266004

Contact: Pierre Genevès

Partner: CHU Grenoble

5.1.2 MuIR

Name: Mu Intermediate Representation

Keywords: Optimizing compiler, Querying

Functional Description: This is a prototype of an intermediate language representation, i.e. an implementation of algebraic terms, rewrite rules, query plans, cost model, query optimizer, and query evaluators. This includes a distributed evaluator of algebraic terms using Apache Spark. Concepts of this implementation have been described in the SIGMOD'20 and CIKM'20 publications, among others, and the distributed evaluator and query optimizers are described in 2021 preprints.

Publications: hal-01673025, hal-03295445, hal-03004218, hal-03517826

Contact: Pierre Genevès

6 New results

6.1 Algebraic Foundations for Distributed Query Evaluation

Participants: Pierre Genevès, Nabil Layaïda, Nils Gesbert, Sarah Chlyah, Amela Fejza, Muideen Lawal, Ugo Comignani, Hadi Dayekh.

An Algebra with a Fixpoint Operator for Distributed Data Collections. Big data programming frameworks are becoming increasingly important for the development of applications, for which performance and scalability are critical. In those complex frameworks, optimizing code by hand is hard and time-consuming, making automated optimization particularly necessary. In order to automate optimization, a prerequisite is to find suitable abstractions to represent programs; for instance, algebras based on monads or monoids to represent distributed data collections. Currently, however, such algebras do not represent recursive programs in a way which allows analyzing or rewriting them. In this paper, we extend a monoid algebra with a fixpoint operator for representing recursion as a first class citizen and show how it allows new optimizations. The fixpoint operator is suitable for modeling recursive computations with distributed data collections. We show that under reasonable conditions this fixpoint can be evaluated by parallel loops with one final merge rather than by a global loop requiring network overhead after each iteration. We also propose several rewrite rules, showing when and how filters can be pushed through recursive terms, and how to filter inside a fixpoint before a join. Experiments with the Spark platform illustrate performance gains brought by these systematic optimizations [11].

Distributed Evaluation of Graph Queries using Recursive Relational Algebra. We have investigated the distributed evaluation of μ-RA queries. We present a system called Dist-μ-RA for the distributed evaluation of recursive graph queries. Dist-μ-RA builds on the recursive relational algebra and extends it with evaluation plans suited for the distributed setting. The goal is to offer expressivity for high-level queries while providing efficiency at scale and reducing communication costs. Experimental results on both real and synthetic graphs show the effectiveness of the proposed approach compared to existing systems [10].

6.2 Exploring Property Graphs with Recursive Path Patterns

We demonstrate a system for recursive query answering over property graphs. The novelty of the system resides in its ability to optimize and efficiently answer recursive path patterns in queries for property graphs. The system is based on a complete implementation of the μ-recursive relational algebra [1]. It also includes parsers and compilers adapted for property graphs so that one can formulate, optimize and answer queries that navigate recursively along paths in property graphs. We demonstrate the system on three real datasets, including the exploration of chains of drug interactions [12].

6.3 Predictive Analytics for Healthcare

Participants: Pierre Genevès, Raouf Kerkouche, Thomas Calmant, Amela Fejza, Luisa Werner, Nabil Layaïda.

Constrained Differentially Private Federated Learning for Low-bandwidth Devices. Federated learning becomes a prominent approach when different entities want to learn collaboratively a common model without sharing their training data. However, Federated learning has two main drawbacks. First, it is quite bandwidth inefficient as it involves a lot of message exchanges between the aggregating server and the participating entities. This bandwidth and corresponding processing costs could be prohibitive if the participating entities are, for example, mobile devices. Furthermore, although federated learning improves privacy by not sharing data, recent attacks have shown that it still leaks information about the training data. This paper presents a novel privacy-preserving federated learning scheme. The proposed scheme provides theoretical privacy guarantees, as it is based on Differential Privacy. Furthermore, it optimizes the model accuracy by constraining the model learning phase on few selected weights. Finally, as shown experimentally, it reduces the upstream and downstream bandwidth by up to 99.9% compared to standard federated learning, making it practical for mobile systems.

These results have been presented at the UAI 2021 conference [7].

Privacy-Preserving and Bandwidth-Efficient Federated Learning: An Application to In-Hospital Mortality Prediction. Machine Learning, and in particular Federated Machine Learning, opens new perspectives in terms of medical research and patient care. Although Federated Machine Learning improves over centralized Machine Learning in terms of privacy, it does not provide provable privacy guarantees. Furthermore, Federated Machine Learning is quite expensive in term of bandwidth consumption as it requires participant nodes to regularly exchange large updates. This paper proposes a bandwidth-efficient privacy-preserving Federated Learning that provides theoretical privacy guarantees based on Differential Privacy. We experimentally evaluate our proposal for in-hospital mortality prediction using a real dataset, containing Electronic Health Records of about one million patients. Our results suggest that strong and provable patient-level privacy can be enforced at the expense of only a moderate loss of prediction accuracy.

These results have been presented at the CHIL 2021 conference [5].

Compression Boosts Differentially Private Federated Learning. Federated Learning allows distributed entities to train a common model collaboratively without sharing their own data. Although it prevents data collection and aggregation by exchanging only parameter updates, it remains vulnerable to various inference and reconstruction attacks where a malicious entity can learn private information about the participants' training data from the captured gradients. Differential Privacy is used to obtain theoretically sound privacy guarantees against such inference attacks by noising the exchanged update vectors. However, the added noise is proportional to the model size which can be very large with modern neural networks. This can result in poor model quality. In this paper, compressive sensing is used to reduce the model size and hence increase model quality without sacrificing privacy. We show experimentally, using 2 datasets, that our privacy-preserving proposal can reduce the communication costs by up to 95% with only a negligible performance penalty compared to traditional non-private federated learning schemes. These results have been presented at the EuroS&P 2021 conference [6].

Scalable and Interpretable Predictive Models for Electronic Health Records. Early identification of patients at risk of developing complications during their hospital stay is currently a challenging issue in healthcare. Complications include hospital-acquired infections, admissions to intensive care units, and in-hospital mortality. Being able to accurately predict the patients' outcomes is a crucial prerequisite for tailoring the care that certain patients receive, if it is believed that they will do poorly without additional intervention. We consider the problem of complication risk prediction, such as inpatient mortality, from the electronic health records of the patients. We study the question of making predictions on the first day at the hospital, and of making updated mortality predictions day after day during the patient's stay. We develop distributed models that are scalable and interpretable. Key insights include analysing diagnoses known at admission and drugs served, which evolve during the hospital stay. We leverage a distributed architecture to learn interpretable models from training datasets of gigantic size. We test our analyses with more than one million of patients from hundreds of hospitals, and report on the lessons learned from these experiments.

Results presented at the 2018 International Conference on Data Science and Applications have been extended with a calibration study and measures for general and instance-level interpretations of the predictions [13].

6.4 Data Cleaning and Exchange

Participants: Ugo Comignani, Angela Bonifati.

Provenance-aware Discovery of Functional Dependencies on Integrated Views. The automatic discovery of functional dependencies(FDs) has been widely studied as one of the hardest problems in data profiling. Existing approaches have focused on making the FD computation efficient while inspecting single relations at a time. In this paper, for the first time we address the problem of inferring FDs for multiple relations as they occur in integrated views by solely using the functional dependencies of the

base relations of the view itself. To this purpose, we leverage logical inference and selective mining and show that we can discover most of the exact FDs from the base relations and avoid the full computation of the FDs for the integrated view itself, while at the same time preserving the lineage of FDs of base relations. We propose algorithms to speedup the inferred FD discovery process and mine FDs on-the-fly only from necessary data partitions. We present InFine(INferred FunctIoNal dEpendency), an end-to-end solution to discover inferred FDs on integrated views by leveraging provenance information of base relations. Our experiments on a range of real-world and synthetic datasets demonstrate the benefits of our method over existing FD discovery methods that need to rerun the discovery process on the view from scratch and cannot exploit lineage information on the FDs. We show that InFine outperforms traditional methods necessitating the full integrated view computation by one to two order of magnitude in terms of runtime. It is also the most memory efficient method while preserving FD provenance information using mainly inference from base table with negligible execution time.

These results will be presented at the ICDE 2022 conference [4].

Explaining Automated Data Cleaning with CLeanEX. We study the explainability of automated data cleaning pipelines and propose CLeanEX, a solution that can generate explanations for the pipelines automatically selected by an automated cleaning system, given it can provide its corresponding cleaning pipeline search space. We propose meaningful explanatory features that are used to describe the pipelines and generate predicate-based explanation rules. We compute quality indicators for these explanations and propose a multi-objective optimization algorithm to select the optimal set of explanations for user-defined objectives. Preliminary experiments show the need for multi-objective optimization for the generation of high-quality explanations that can be either intrinsic to the single selected cleaning pipeline or relative to the other pipelines that were not selected by the automated cleaning system. We also show that CLeanEX is a promising step towards generating automatically insightful explanations, while catering to the needs of the user alike. [2].

Exchanging Data under Policy Views. Exchanging data between data sources is a fundamental problem in many data science and data integration tasks. In this paper, we focus on the data exchange problem in the presence of privacy constraints on the source data, which has been disregarded in the literature to date. By leveraging a logical privacy-preservation paradigm, the privacy restrictions are expressed as a set of policy views representing the information that is safe to expose over all instances of the source in order to exchange them with the target. We introduce a protocol that provides formal privacy guarantees and is data-independent, i.e., under certain criteria, it guarantees that the mappings leak no sensitive information independently of the instances lying in the source. Moreover, we design an algorithm for repairing an input mapping w.r.t. a set of policy views, in cases where the input mapping leaks sensitive information. We show that the repairing can build upon hard-coded and learning-based user preference functions and we show the trade-offs. Our empirical evaluation shows that repairing mappings is quite efficient, leading to repairing sets of 300 s-t tgds in an average time of 5s on a commodity machine. It also shows that the repairing based on learning is robust and has comparable runtimes with the hard-coded one [3].

7 Partnerships and cooperations

7.1 National initiatives

7.1.1 ANR

CLEAR

Participants: Pierre Genevès, Nabil Layaïda, Nils Gesbert, Sarah Chlyah, Muideen Lawal, Amela Fejza, Thomas Calmant, Hadi Dayekh.

- Title: Compilation of intermediate Languages into Efficient big dAta Runtimes

- Call: Appel à projets générique 2016 défi 'Société de l'information et de la communication' – JCJC
- Duration: January 2017 – Mars 2022
- Coordinator: Pierre Genevès
- See also: tyrex.inria.fr/clear
- Abstract: This project addresses one fundamental challenge of our time: the construction of effective programming models and compilation techniques for the correct and efficient exploitation of big and linked data. We study high-level specifications of pipelines of data transformations and extraction for producing valuable knowledge from rich and heterogeneous data. We investigate how to synthesize code which is correct and optimized for execution on distributed infrastructures.

QualiHealth

Participants: Ugo Comignani, Angela Bonifati.

- Title: Enhancing the Quality of Health Data
- Call: Appel à projets Projets de Recherche Collaborative – Entreprise (PRCE)
- Duration: 2018-2022
- Coordinator: Angela Bonifati
- Others partners: LIMOS, Université Clermont Auvergne. LIS, Université d'Aix-Marseille. HEGP, INSERM, Paris. Inst. Cochin, INSERM, Paris. Gnubila, Argonay. The University of British Columbia, Vancouver (Canada)
- Abstract: This research project is geared towards a system capable of capturing and formalizing the knowledge of data quality from domain experts, enriching the available data with this knowledge and thus exploiting this knowledge in the subsequent quality-aware medical research studies. We expect a quality-certified collection of medical and biological datasets, on which quality-certified analytical queries can be formulated. We envision the conception and implementation of a quality-aware query engine with query enrichment and answering capabilities.

  To reach this ambitious objectives, the following concrete scientific goals must be fulfilled : (1) An innovative research approach, that starts from concrete datasets and expert practices and knowledge to reach formal models and theoretical solutions, will be employed to elicit innovative quality dimensions and to identify, formalize, verify and finally construct quality indicators able to capture the variety and complexity of medical data; those indicators have to be composed, normalized and aggregated when queries involve data with different granularities (e.g., accuracy indications on pieces of information at the patient level have to be composed when one queries cohort) and of different quality dimensions (e.g., mixing incomplete and inaccurate data); and (2) In turn, those complex aggregated indicators have to be used to provide new quality-driven query answering, refinement, enrichment and data analytics techniques. A key novelty of this project is the handling of data which are not rectified on the original database but sanitized in a query-driven fashion: queries will be modified, rewritten and extended to integrate quality parameters in a flexible and automatic way.

7.2 Regional initiatives

- P. Genevès is member of the board of the Deepcare MIAI Chair, led by Philippe Cinquin.
- N. Layaïda and P. Genevès are members of the MIAI Knowledge communication and evolution chair, led by Jérôme Euzenat.

8 Dissemination

Participants: Pierre Genevès, Nabil Layaïda, Nils Gesbert, Ugo Comignani, Angela Bonifati.

8.1 Promoting scientific activities

8.1.1 Scientific events: selection

Member of the conference program committees

- P. Genevès is PC member for the ACM SIGMOD 2023 conference.
- U. Comignani is PC member for the ACM SIGMOD 2022 conference.

8.1.2 Journal

Reviewer - reviewing activities

- U. Comignani is reviewer for the VLDB journal.

8.1.3 Invited talks

- P. Genevès was an invited speaker at the Inria-ATOS workshop in 2021.

8.1.4 Research administration

- P. Genevès is member of the board of the CNRS LIG laboratory, responsible for the "formal methods models and languages" axis of the laboratory.
- P. Genevès is co-responsible of the Doctoral School MSTII, responsible for the Computer Science specialty.
- N. Layaïda is a member of the experts pool (selection committee) of the minalogic competitive cluster.
- N. Layaïda is a member of the scientific comittee of the LabEx PERSYVAL-lab (Pervasive Systems and Algorithms).
- N. Layaïda is a member of the Scientific Board of Digital League, the digital cluster of Auvergne-Rhône-Alpes.

8.2 Teaching - Supervision - Juries

8.2.1 Teaching

- P. Genevès is co-responsible of the M2-level course ‘Fundamentals of Data Processing and Distributed Knowledge’ of the MOSIG program at UGA (36h)
- P. Genevès is co-responsible of the M2-level course ‘Accès à l’information: du web des données au web sémantique’ in the ENSIMAG ISI 3A program at Grenoble-INP (30h)
- Master : N. Gesbert, academic tutorship of an apprentice, 10 h eq TD, M1, Grenoble INP
- Master : N. Gesbert, ‘Construction d’applications Web’, 42 h eq TD, M1, Grenoble INP
- Master : N. Gesbert, ‘Principes des systèmes de gestion des bases de données’, 36 h eq TD, M1, Grenoble INP

- Master : N. Gesbert, ‘Introduction to lambda-calculus’, 7 h 30 eq TD, M2, UGA-Grenoble INP (MOSIG)
- Licence : N. Gesbert, ‘Logique pour l’informatique’, 45 h eq TD, L3, Grenoble INP
- N. Gesbert is in charge of the L3-level course ‘logique pour l’informatique’ and of the M1-level course ‘Principes des systèmes de gestion de bases de données (SEOC)’.
- Master : U. Comignani is co-responsible of the "BigData" master, co-accredited between Grenoble Ecole de Management and Grenoble INP
- Master : U. Comignani is in charge of the ’Projets fil rouges’, 10 h eq TD, MS BigData, Grenoble INP
- Master : U. Comignani, ‘Principes des systèmes de gestion de bases de données’, 90 h eq TD, M1, Grenoble INP
- Master : U. Comignani, ‘Projet BD’, 15 h eq TD, M1, Grenoble INP
- Master : U. Comignani, ‘Stockage et traitement de données à grande échelle’, 12 h eq TD, M2, Grenoble INP
- Master : U. Comignani, academic tutorship of an apprentice, 10 h eq TD, M1, Grenoble INP
- A. Bonifati taught a course on Graph-based Knowledge Representation at in the Master Informatique Fondamentale (M2) at ENS Lyon.

8.2.2 Supervision

- PhD defended in April 2021 by Muideen Lawal on ’On Cost Estimation for the Recursive Relational Algebra’, PhD started in October 2017, co-supervised by Pierre Genevès and Nabil Layaïda [9].
- PhD defended in July 2021 by Raouf Kerkouche on ’Differentially Private Federated Learning for Bandwidth and Energy Constrained Environments’, PhD started in October 2017, co-supervised by Pierre Genevès and Claude Castelluccia [8].
- PhD in progress: Sarah Chlyah, Algebraic foundations for the synthesis of optimized distributed code, PhD started in March 2018, co-supervised by Pierre Genevès, Nils Gesbert and Nabil Layaïda.
- PhD in progress: Amela Fejza, On the extended algebraic representations for analytical workloads, PhD started in October 2018, supervised by Pierre Genevès.
- PhD in progress: Luisa Werner, Neural Symbolic Integration, PhD started in October 2020, co-supervised by Nabil Layaïda and Pierre Genevès.
- M2 Internship of Hadi Dayekh defended in June 2021, co-supervised by Ugo Comignani and Pierre Genevès

8.2.3 Juries

- P. Genevès has been jury president for the PhD thesis of Line Van Den Berg, entitled ‘Cultural knowledge evolution in dynamic epistemic logic’. PhD thesis of University Grenoble Alpes defended on October 29th, 2021.
- N. Layaïda has been jury member and rapporteur of Simon Pierre Dembele thesis entitled ’Auditer l’Énergie – Avant de Déployer ses Modèles : Vers des optimiseurs verts de requêtes analytiques’, Thesis of Ecole Nationale Supérieure de Mécanique et d’Aérotechnique, Poitiers, defended on the 8th July, 2021.

8.3 Popularization

8.3.1 Articles and contents

Angela Bonifati co-published an article in CACM [14] which is a community view, as a follow-up to the Dagstuhl Seminar 19491 on Big Graph Processing Systems that she co-organized:

The Future is Big Graphs! A Community View on Graph Processing Systems. Graphs are by nature unifying abstractions that can leverage interconnectedness to represent, explore, predict, and explain real- and digital-world phenomena. Although real users and consumers of graph instances and graph workloads understand these abstractions, future problems will require new abstractions and systems. What needs to happen in the next decade for big graph processing to continue to succeed? This is a view published in CACM [14].

9 Scientific production

9.1 Major publications

[1] L. Jachiet, P. Genevès, N. Gesbert and N. Layaïda. 'On the Optimization of Recursive Relational Queries: Application to Graph Queries'. In: *SIGMOD 2020 - ACM International Conference on Management of Data*. Portland, United States, June 2020, pp. 1–23. DOI: 10.1145/3318464.3380567. URL: https://hal.inria.fr/hal-01673025.

9.2 Publications of the year

International peer-reviewed conferences

[2] L. Berti-Équille and U. Comignani. 'Explaining Automated Data Cleaning with CLeanEX'. In: IJCAI-PRICAI 2020 - Workshop on Explainable Artificial Intelligence (XAI). Online, Japan, 8th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03148996.

[3] A. Bonifati, U. Comignani and E. Tsamoura. 'Exchanging Data under Policy Views'. In: EDBT 2021 - 24th International Conference on Extending Database Technology. Nicosia, Cyprus, 23rd Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03149043.

[4] U. Comignani, L. Berti-Équille, N. Novelli and A. Bonifati. 'Provenance-aware Discovery of Functional Dependencies on Integrated Views'. In: ICDE 2022 - 38th IEEE International Conference on Data Engineering. Kuala Lumpur, Malaysia: IEEE, 9th May 2022, pp. 1–15. URL: https://hal.archives-ouvertes.fr/hal-03474955.

[5] R. Kerkouche, G. Acs, C. Castelluccia and P. Genevès. 'Privacy-Preserving and Bandwidth-Efficient Federated Learning: An Application to In-Hospital Mortality Prediction'. In: CHIL 2021 - ACM Conference on Health, Inference, and Learning. virtual event, France: ACM, 8th Apr. 2021, pp. 1–11. URL: https://hal.inria.fr/hal-03160473.

[6] R. Kerkouche, G. Ács, C. Castelluccia and P. Genevès. 'Compression Boosts Differentially Private Federated Learning'. In: EuroS&P 2021 - 6th IEEE European Symposium on Security and Privacy. Vienna, Austria: IEEE, 6th Sept. 2021, pp. 1–15. URL: https://hal.archives-ouvertes.fr/hal-03066941.

[7] R. Kerkouche, G. Ács, C. Castelluccia and P. Genevès. 'Constrained Differentially Private Federated Learning for Low-bandwidth Devices'. In: *Proceedings of Machine Learning Research*. UAI 2021 - 37th Conference on Uncertainty in Artificial Intelligence. Online, United States, 26th July 2021, pp. 1–18. URL: https://hal.archives-ouvertes.fr/hal-03266004.

Doctoral dissertations and habilitation theses

[8] R. Kerkouche. 'Differentially Private Federated Learning for Bandwidth and Energy Constrained Environments'. UGA, 7th July 2021. URL: https://hal.inria.fr/tel-03551408.

[9] M. Lawal. 'On Cost Estimation for the Recursive Relational Algebra'. UGA, 21st Apr. 2021. URL: https://hal.inria.fr/tel-03551396.

Reports & preprints

[10] S. Chlyah, P. Genevès and N. Layaïda. *Distributed Evaluation of Graph Queries using Recursive Relational Algebra*. 24th Nov. 2021. URL: https://hal.inria.fr/hal-03295445.

[11] S. Chlyah, N. Gesbert, P. Genevès and N. Layaïda. *On the Optimization of Iterative Programming with Distributed Data Collections*. 2nd Mar. 2021. URL: https://hal.inria.fr/hal-02066649.

[12] A. Fejza, P. Genevès and N. Layaïda. *Exploring Property Graphs with Recursive Path Patterns*. 10th Jan. 2022. URL: https://hal.inria.fr/hal-03517826.

[13] A. Fejza, P. Genevès, N. Layaïda and J.-L. Bosson. *Scalable and Interpretable Predictive Models for Electronic Health Records*. 29th Jan. 2021. URL: https://hal.inria.fr/hal-03124966.

9.3 Other

Scientific popularization

[14] S. Sakr, A. Bonifati, H. Voigt, A. Iosup, K. Ammar, R. Angles, W. Aref, M. Arenas, M. Besta, P. A. Boncz, K. Daudjee, E. Della Valle, S. Dumbrava, O. Hartig, B. Haslhofer, T. Hegeman, J. Hidders, K. Hose, A. Iamnitchi, V. Kalavri, H. Kapp, W. Martens, T. Özsu, E. Peukert, S. Plantikow, M. Ragab, M. R. Ripeanu, S. Salihoglu, C. Schulz, P. Selmer, J. F. Sequeda, J. Shinavier, G. Szárnyas, R. Tommasini, A. Tumeo, A. Uta, A. L. Varbanescu, H.-Y. Wu, N. Yakovets, D. Yan and E. Yoneki. 'The Future is Big Graphs! A Community View on Graph Processing Systems'. In: *Communications of the ACM* 64.9 (Sept. 2021), pp. 62–71. DOI: 10.1145/3434642. URL: https://hal.inria.fr/hal-03128601.