RESEARCH CENTRE

Grenoble - Rhône-Alpes

IN PARTNERSHIP WITH:

CNRS, Institut polytechnique de Grenoble, Université de Grenoble Alpes

2020
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 Team: 2012 November 01, updated into 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

B9.11.2. – Financial risks

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, HDR]
- Ugo Comignani [Institut polytechnique de Grenoble, Associate Professor, from Sep 2020]
- Nils Gesbert [Institut polytechnique de Grenoble, Associate Professor]
- Cécile Roisin [Univ Grenoble Alpes, Professor, until Sep 2020, HDR]

PhD Students

- Fateh Boulmaiz [Milev Multiservice Sas]
- Sarah Chlyah [Inria]
- Amela Fejza [Univ Grenoble Alpes]
- Muideen Lawal [Univ Grenoble Alpes]
- Luisa Werner [Univ Grenoble Alpes, from Oct 2020]

Technical Staff

- Thomas Calmant [Inria, Engineer]

Interns and Apprentices

- Luisa Werner [Institut de technologie de Karlsruhe - Allemagne, until Apr 2020]

Administrative Assistant

- Helen Pouchot-Rouge-Blanc [Inria]

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 Highlights of the year

We obtained fundamental results on the optimization of recursive relational queries. We extended the relational algebra with the support of recursive terms and with new algebraic transformation rules. These extensions make it possible to compute query evaluation plans that were not reachable with earlier approaches. In practice, this translates into drastic performance gains for the evaluation of recursive queries over graphs. A part of these results were presented at the SIGMOD 2020 conference [1].

6 New software and platforms

6.1 New software

6.1.1 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. This prototype implementation is described in the 2017 preprint available at: https://hal.inria.fr/hal-01517087/document.

Publication: hal-01517087

Contact: Pierre Genevès

Participants: Pierre Genevès, Thomas Calmant

Partner: CHU Grenoble

6.1.2 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 (including a distributed evaluator of algebraic terms using Apache Spark).

Contact: Pierre Genevès

7 New results

7.1 On the Optimization of Recursive Relational Queries: Application to Graph Queries

Graph databases have received a lot of attention as they are particularly useful in many applications such as social networks, life sciences and the semantic web. Various languages have emerged to query graph databases, many of which embed forms of recursion which reveal essential for navigating in graphs. The relational model has benefited from a huge body of research in the last half century and that is why many graph databases rely on techniques of relational query engines. Since its introduction, the relational model has seen various attempts to extend it with recursion and it is now possible to use recursion in several SQL or Datalog based database systems. The optimization of recursive queries remains, however, a challenge. We propose μ-RA, a variation of the Relational Algebra equipped with a fixpoint operator for expressing recursive relational queries. μ-RA can notably express unions of conjunctive regular path queries. Leveraging the fact that this fixpoint operator makes recursive terms more amenable to algebraic transformations, we propose new rewrite rules. These rules makes it possible to generate new query execution plans, that cannot be obtained with previous approaches. We present the syntax and semantics of μ-RA, and the rewriting rules that we specifically devised to tackle the optimization of recursive queries.

We report on practical experiments that show that the newly generated plans can provide significant performance improvements for evaluating recursive queries over graphs.

These results have been presented at the SIGMOD 2020 conference [9].

One main advantage of the approach is to make it possible to compute new query evaluation plans that were not reachable with previous approaches. For selecting an efficient evaluation plan among millions of variants, we rely on cost estimations. Results on the cost estimation technique for recursive terms were presented at the CIKM 2020 conference [12].

7.2 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 [18].

7.3 Backward Type Inference for XML Queries

Although XQuery is a statically typed, functional query language for XML data, some of its features such as upward and horizontal XPath axes are typed imprecisely. The main reason is that while the XQuery data model allows to navigate upwards and between siblings from a given XML node, the type model, e.g., regular tree types, can only describe the subtree structure of the given node. In 2015, Giuseppe Castagna and our team independently proposed a precise forward type inference system for XQuery using an extended type language that can describe not only a given XML node but also its context. Recently, as a complementary method to such forward type inference systems, we propose an enhanced backward type inference system for XQuery, based on an extended type language. Results include an exact type system for XPath axes and a sound type system for XQuery expressions.

These results have been published in the journal Theoretical Computer Science [5]

7.4 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 [19].

7.5 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 will be presented at the EuroS&P 2021 conference [11].

7.6 Predictive Analytics by Inferring Structure from Electronic Health Records

Data science in the domain of healthcare recently registered considerable results in predicting medical outcomes based on Electronic Health Records. Especially deep learning methods enhance prediction performance by finding meaningful representations of input data. Latest studies show that Electronic Health Records explicitly or implicitly contain underlying causal relations which can be modeled as directed acyclic graphs. Recently, Graph Convolutional Transformer was proposed to learn the implicit graph structure of Electronic Health Records based on a method called attention. Then, Graph Convolutional Transformer exploits this structure to extract representations and conduct prediction tasks on Electronic Health Records. In this work, Graph Convolutional Transformer is applied on a large data set from Premier Healthcare Database in order to perform mortality prediction for patients who are admitted to hospitals. This work shows that Graph Convolutional Transformer leads to a state-of-the-art performance on the mortality prediction task in comparison to the applied baseline models. Furthermore, some possible extensions of Graph Convolutional Transformer are illustrated that have the potential to further improve the predictive performance of Graph Convolutional Transformer.

This has been the topic of a Master's thesis [23].

7.7 Graph Generators: State of the Art and Open Challenges

We focus on the computational complexity of regular simple path queries (RSPQs). We consider the following problem RSPQ(L) for a regular language L: given an edge-labeled digraph Gand two nodes xand y, is there a simple path from x to y that forms a word belonging to L? We fully characterize the frontier between tractability and intractability for RSPQ(L). More precisely, we prove RSPQ(L)is either AC0, NL-complete or NP-complete depending on the language L. We also provide a simple characterization of the tractable fragment in terms of regular expressions. Finally, we also discuss the complexity of deciding whether a language L belongs to the fragment above. We consider several alternative representations of L: DFAs, NFAs or regular expressions, and prove that this problem is NL-complete for the first representation and PSpace-complete for the other two.

These works were published in the ACM Computing Surveys journal [3].

7.8 Evaluating Top-k Queries with Inconsistency Degrees

We study the problem of augmenting relational tuples with inconsistency awareness and tackling top-k queries under a set of denial constraints (DCs). We define a notion of inconsistent tuples with respect to a set of DCs and define two measures of inconsistency degrees, which consider single and multiple violations of constraints. In order to compute these measures, we leverage two models of provenance, namely why-provenance and provenance polynomials. We investigate top-k queries that allow to rank the answer tuples by their inconsistency degrees. Since one of our measure is monotonic and the other non-monotonic, we design an integrated top-k algorithm to compute the top-k results of a query w.r.t. both inconsistency measures. By means of an extensive experimental study, we gauge the effectiveness of inconsistency-aware query answering and the efficiency of our algorithm with respect to a baseline, where query results are fully computed and ranked afterwards.

These results were published in the VLDB journal [6].

7.9 An Analytical Study of Large SPARQL Query Logs

With the adoption of RDF as the data model for Linked Data and the Semantic Web, query specification from end-users has become more and more common in SPARQL endpoints. In this paper, we conduct an in-depth analytical study of the queries formulated by end-users and harvested from large and up-to-date structured query logs from a wide variety of RDF datasources. As opposed to previous studies, ours is the first assessment on a voluminous query corpus, spanning over several years and covering many representative SPARQL endpoints. Apart from the syntactical structure of the queries, that exhibits already interesting results on this generalized corpus, we drill deeper in the structural characteristics related to the graph and hypergraph representation of queries. We outline the most common shapes of queries when visually displayed as undirected graphs, characterize their tree width, length of their cycles, maximal degree of nodes, and more. For queries that cannot be adequately represented as graphs, we investigate their hypergraphs and hypertree width. Moreover, we analyze the evolution of queries over time, by introducing the novel concept of a streak, i.e., a sequence of queries that appear as subsequent modifications of a seed query. Our study offers several fresh insights on the already rich query features of real SPARQL queries formulated by real users, and brings us to draw a number of conclusions and pinpoint future directions for SPARQL query evaluation, query optimization, tuning, and benchmarking.

These results were published in the VLDB journal [4].

7.10 A relational framework for inconsistency-aware query answering

We introduce a novel framework for encoding inconsistency into relational tuples and tackling query answering for union of con-junctive queries (UCQs) with respect to a set of denial constraints (DCs). We define a notion of inconsistent tuple with respect to a set of DCs and define four measures of inconsistency degree of an answer tuple of a query. Two of these measures revolve around the minimal number of inconsistent tuples necessary to compute the answer tuples of a UCQ, whereas the other two rely on the maximum number of inconsistent tuples under set-and bag-semantics, respectively. In order to compute these measures of inconsistency degree, we leverage two models of provenance semiring, namely why-provenance and provenance polynomials, which can be computed in polynomial time in the size of the relational instances for UCQs. Hence, these measures of inconsistency degree are also computable in polynomial time in data complexity. We also investigate top-k and bounded query answering by ranking the answer tuples by their inconsistency degrees. We explore both a full materialized approach and a semi-materialized approach for the computation of top-k and bounded query results [20].

7.11 A trichotomy for regular simple path queries on graphs

We focus on the computational complexity of regular simple path queries (RSPQs). We consider the following problem RSPQ(L) for a regular language L: given an edge-labeled digraph G and two nodes x and y, is there a simple path from x to y that forms a word belonging to L? We fully characterize the frontier between tractability and intractability for RSPQ(L). More precisely, we prove RSPQ(L) is either AC0, NL-complete or NP-complete depending on the language L. We also provide a simple characterization of the tractable fragment in terms of regular expressions. Finally, we also discuss the complexity of deciding whether a language L belongs to the fragment above. We consider several alternative representations of L: DFAs, NFAs or regular expressions, and prove that this problem is NL-complete for the first representation and PSpace-complete for the other two.

These results were published in the JCSS journal [2].

7.12 RDF graph anonymization robust to data linkage

Privacy is a major concern when publishing new datasets in the context of Linked Open Data (LOD). A new dataset published in the LOD is indeed exposed to privacy breaches due to the linkage to objects already present in the other datasets of the LOD. In this paper, we focus on the problem of building safe anonymizations of an RDF graph to guarantee that linking the anonymized graph with any external RDF graph will not cause privacy breaches. Given a set of privacy queries as input, we study the data-independent safety problem and the sequence of anonymization operations necessary to enforce it. We provide sufficient conditions under which an anonymization instance is safe given a set of privacy queries. Additionally, we show that our algorithms for RDF data anonymization are robust in the presence of "`sameAs`" links that can be explicit or inferred by additional knowledge. These works were published in the WISE 2019 congress that finally took place in 2020 due to events in Hong-Kong [8].

7.13 Regular Path Query Evaluation on Streaming Graphs

We study persistent query evaluation over streaming graphs, which is becoming increasingly important. We focus on navigational queries that determine if there exists a path between two entities that satisfies a user-specified constraint. We adopt the Regular Path Query (RPQ) model that specifies navigational patterns with labeled constraints. We propose deterministic algorithms to efficiently evaluate persistent RPQs under both arbitrary and simple path semantics in a uniform manner. Experimental analysis on real and synthetic streaming graphs shows that the proposed algorithms can process up to tens of thousands of edges per second and efficiently answer RPQs that are commonly used in real-world workloads.

These results were presented in the SIGMOD conference [13].

7.14 SHARQL: Shape Analysis of Recursive SPARQL Queries

We showcase SHARQL, a system that allows to navigate SPARQL query logs, can inspect complex queries by visualizing their shape, and can serve as a back-end to flexibly produce statistics about the logs. Even though SPARQL query logs are increasingly available and have become public recently, their navigation and analysis is hampered by the lack of appropriate tools. SPARQL queries are sometimes hard to understand and their inherent properties, such as their shape, their hypertree properties, and their property paths are even more difficult to be identified and properly rendered. In SHARQL, we show how the analysis and exploration of several hundred million queries is possible. We offer edge rendering which works with complex hyperedges, regular edges, and property paths of SPARQL queries. The underlying database stores more than one hundred attributes per query and is therefore extremely flexible for exploring the query logs and as a back-end to compute and display analytical properties of the entire logs or parts thereof.

These results were presented in the SIGMOD conference [7].

7.15 Graph Summarization

The continuous and rapid growth of highly interconnected datasets, which are both voluminous and complex, calls for the development of adequate processing and analytical techniques. One method for condensing and simplifying such datasets is graph summarization. It denotes a series of application-specific algorithms designed to transform graphs into more compact representations while preserving structural patterns, query answers, or specific property distributions. As this problem is common to several areas studying graph topologies, different approaches, such as clustering, compression, sampling, or influence detection, have been proposed, primarily based on statistical and optimization methods. The focus of our chapter is to pinpoint the main graph summarization methods, but especially to focus on the most recent approaches and novel research trends on this topic, not yet covered by previous surveys [17].

7.16 Valentine: Evaluating Matching Techniques for Dataset Discovery

Data scientists today search large data lakes to discover and integrate datasets. In order to bring together disparate data sources, dataset discovery methods rely on some form of schema matching: the process of establishing correspondences between datasets. Traditionally, schema matching has been used to find matching pairs of columns between a source and a target schema. However, the use of schema matching in dataset discovery methods differs from its original use. Nowadays schema matching serves as a building block for indicating and ranking inter-dataset relationships. Surprisingly, although a discovery method's success relies highly on the quality of the underlying matching algorithms, the latest discovery methods employ existing schema matching algorithms in an ad-hoc fashion due to the lack of openly-available datasets with ground truth, reference method implementations, and evaluation metrics. In this paper, we aim to rectify the problem of evaluating the effectiveness and efficiency of schema matching methods for the specific needs of dataset discovery. To this end, we propose Valentine, an extensible open-source experiment suite to execute and organize large-scale automated matching experiments on tabular data. Valentine includes implementations of seminal schema matching methods that we either implemented from scratch (due to absence of open source code) or imported from open repositories. The contributions of Valentine are: i) the definition of four schema matching scenarios as encountered in dataset discovery methods, ii) a principled dataset fabrication process tailored to the scope of dataset discovery methods and iii) the most comprehensive evaluation of schema matching techniques to date, offering insight on the strengths and weaknesses of existing techniques, that can serve as a guide for employing schema matching in future dataset discovery methods [21].

7.17 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. [14].

7.18 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 [15].

7.19 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 [22].

8 Partnerships and cooperations

8.1 National initiatives

8.1.1 ANR

CLEAR

- 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: http://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.

DataCert

- Title: Coq deep specification of security aware data integration
- Call: Appel à projets Sciences et technologies pour la confiance et la sécurité numérique
- Duration: January 2016 – January 2020
- Participant: Angela Bonifati
- Others partners: Université Paris Sud/Laboratoire de Recherche en Informatique, Université de Lille/Centre de Recherche en Informatique, Signal et Automatique de Lille, Université de Lyon/Laboratoire d'InfoRmatique en Image et Systèmes d'information.
- See also: http://datacert.lri.fr/
- Abstract: This project's aim is to develop a comprehensive framework handling the fundamental problems underlying security-aware data integration and sharing, resulting in a paradigm shift in the design and implementation of security-aware data integration systems. To fill the gap between both worlds, we strongly rely on deep specifications and proven-correct software, develop formal models yielding highly reliable technology while controlling the disclosure of private or confidential information.

QualiHealth

- 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.

8.2 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.

9 Dissemination

9.1 Promoting scientific activities

9.1.1 Scientific events: organisation

General chair, scientific chair

- A. Bonifati was Program Chair of EDBT 2020: 23rd International Conference on Extending Database Technology, 30th March-2nd April, 2020. Copenhagen, Denmark.

9.1.2 Scientific events: selection

Chair of conference program committees

- A. Bonifati was Demo Co-Chair of ICDE 2020 and the Sigmod 2020 Workshops.

Member of the conference program committees

- P. Genevès was a Program Committee member of IJCAI 2020, the 29th International Joint Conference on Artificial Intelligence 2020.
- P. Genevès was a Program Committee member of AAAI 2020, the 34h AAAI Conference on Artificial Intelligence 2020.
- A. Bonifati was Conference Program Committee of IEEE BigData 2020.
- A. Bonifati was Conference Program Committee of VLDB 2020.

9.1.3 Journal

Member of the editorial boards

- A. Bonifati is Associate Editor of ACM TODS, The VLDB Journal, Distributed and Parallel Databases and Frontiers in Big Data.

9.1.4 Invited talks

- A. Bonifati was an invited speaker at the 6th International (online) Summer school on AI and Big Data.
- P. Genevès was an invited speaker at the Inria-ATOS workshop in 2020.

9.1.5 Leadership within the scientific community

- A. Bonifati is President of the EDBT Board (since July 2020).
- P. Genevès was member of the committee for the best PhD thesis award of BDA 2020.

9.1.6 Scientific expertise

Members of the project were experts for European Commission (H2020), ANR, NSERC (Canada), SEDA (Latvia).

9.1.7 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 has been vice-president of the Inria CRCN-ISFP hiring committee for Inria Grenoble - Rhône-Alpes research center (58 candidates for 6 positions).
- N. Layaïda has been a member of the hiring committee of an Assistant Professor position at Ensimag, Grenoble INP (ENSIMAG COS 27 MCF 0665).
- N. Layaïda is a member of the experts pool (selection committee) of the minalogic competitive cluster.
- A. Bonifati and N. Layaïda are members of the Scientific Board of Digital League, the digital cluster of Auvergne-Rhône-Alpes.
- N. Layaïda is a member of the scientific comittee of the LabEx PERSYVAL-lab (Pervasive Systems and Algorithms).

9.2 Teaching - Supervision - Juries

9.2.1 Teaching

- Licence : C. Roisin, Programmation C, 12h eq TD, L2, IUT2, Univ. Grenoble-Alpes
- Licence : C. Roisin, Architecture des réseaux, 112h eq TD, L1, IUT2, Univ. Grenoble-Alpes
- Licence : C. Roisin, Services réseaux, 22h eq TD, L2, IUT2, Univ. Grenoble-Alpes
- Licence : C. Roisin, Introduction système Linux, 21h eq TD, L1, IUT2, Univ. Grenoble-Alpes
- Licence : C. Roisin, Système et réseaux, 14h eq TD, L3, IUT2, Univ. Grenoble-Alpes

- Licence : C. Roisin, Tutorat pédagogique de 4 apprentis, 20h eq TD, L3, IUT2, Univ. Grenoble-Alpes
- Licence : C. Roisin, Suivi pédagogique de 20 étudiants (responsable de la Licence Professionnelle MI-ASSR), 13h eq TD, L1, IUT2, Univ. Grenoble-Alpes
- Licence : N. Gesbert, 'Logique pour l'informatique', 45 h eq TD, L3, Grenoble INP
- Master : N. Gesbert, academic tutorship of an apprentice, 10 h eq TD, M1, Grenoble INP
- Master : N. Gesbert, 'Construction d'applications Web', 21 h eq TD, M1, Grenoble INP
- Master : N. Gesbert, 'Analyse, conception et validation de logiciels', 30 h eq TD, M1, Grenoble INP
- Master : N. Gesbert, 'Introduction to lambda-calculus', 5 h eq TD, M2, UGA-Grenoble INP (MOSIG)
- N. Gesbert is responsible of the L3-level course 'logique pour l'informatique' (25 apprentices) and of the M1-level course 'construction d'applications Web' (72 students).
- Master : U. Comignani, 'Principes des systèmes de gestion de bases de données', 60 h eq TD, M1, Grenoble INP
- Master : U. Comignani, 'Projet BD', 37.5 h eq TD, M1, Grenoble INP
- Master : U. Comignani, 'Stockage et traitement de données à grande échelle', 12 h eq TD, M2, Grenoble INP
- P. Genevès is co-responsible and teacher in the M2-level course 'Semantic Web: from XML to OWL' of the MOSIG program at UGA (36h)
- P. Genevès is co-responsible and teacher in the M2-level course 'Accès à l'information: du web des données au web sémantique' of the ENSIMAG ISI 3A program at Grenoble-INP (30h)
- A. Bonifati teached a course on Graph-based Knowledge Representationat in the Master Informatique Fondamentale (M2) at ENS Lyon.

9.2.2 Supervision

- PhD in progress: Muideen Lawal, Cost models for optimizing compilers based on mu-terms, PhD started in October 2017, co-supervised by Pierre Genevès and Nabil Layaïda.
- PhD in progress: Raouf Kerkouche, Privacy-preserving predictive analytics with big prescription data, PhD started in October 2017, co-supervised by Pierre Genevès and Claude Castelluccia.
- 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.

9.3 Popularization

- A. Fejza participated in the development GazePlay project: its aims at developing open games in order to support individuals with multiple disabilities to interact with their environment. It is a free and open-source software which gathers several mini-games playable with all eye-trackers including low cost ones [16].

10 Scientific production

10.1 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.

10.2 Publications of the year

International journals

[2] G. Bagan, A. Bonifati and B. Groz. 'A trichotomy for regular simple path queries on graphs'. In: *Journal of Computer and System Sciences* 108 (Mar. 2020), pp. 29–48. DOI: 10.1016/j.jcss.2019.08.006. URL: https://hal.inria.fr/hal-02435355.

[3] A. Bonifati, I. Holubovà, A. Prat-Pérez and S. Sakr. 'Graph Generators: State of the Art and Open Challenges'. In: *ACM Computing Surveys* 53.2 (Apr. 2020). DOI: 10.1145/3379445. URL: https://hal.inria.fr/hal-02435371.

[4] A. Bonifati, W. Martens and T. Timm. 'An Analytical Study of Large SPARQL Query Logs'. In: *The VLDB Journal* 29.2-3 (1st June 2020). URL: https://hal.archives-ouvertes.fr/hal-03118422.

[5] H. Im, P. Genevès, N. Gesbert and N. Layaïda. 'Backward Type Inference for XML Queries'. In: *Theoretical Computer Science*. Theoretical Computer Science 823 (27th Mar. 2020), pp. 69–99. DOI: 10.1016/j.tcs.2020.03.020. URL: https://hal.inria.fr/hal-01497857.

[6] O. Issa, A. Bonifati and F. Toumani. 'Evaluating Top-k Queries with Inconsistency Degrees'. In: *Proceedings of the VLDB Endowment (PVLDB)* (June 2020). DOI: 10.14778/3407790.3407815. URL: https://hal.archives-ouvertes.fr/hal-02898931.

International peer-reviewed conferences

[7] A. Bonifati, W. Martens and T. Timm. 'SHARQL: Shape Analysis of Recursive SPARQL Queries'. In: SIGMOD/PODS 2020 - International Conference on Management of Data. Portland OR, United States, 14th June 2020, pp. 2701–2704. DOI: 10.1145/3318464.3384684. URL: https://hal.inria.fr/hal-03125718.

[8] R. Delanaux, A. Bonifati, M.-C. Rousset and R. Thion. 'RDF graph anonymization robust to data linkage'. In: WISE 2019 - 20th International Conference on Web Information Systems Engineering. Vol. 11881. Lecture Notes in Computer Science (LNCS). Hong Kong, China, 29th Oct. 2019, pp. 491–506. DOI: 10.1007/978-3-030-34223-4_31. URL: https://hal.archives-ouvertes.fr/hal-02444752.

[9] 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, 14th June 2020, pp. 1–23. DOI: 10.1145/3318464.3380567. URL: https://hal.inria.fr/hal-01673025.

[10] 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, 8th Apr. 2021, pp. 1–11. URL: https://hal.inria.fr/hal-03160473.

[11] 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, 6th Sept. 2021, pp. 1–15. URL: https://hal.archives-ouvertes.fr/hal-03066941.

[12] M. Lawal, P. Genevès and N. Layaïda. ‘A Cost Estimation Technique for Recursive Relational Algebra’. In: CIKM 2020 - 29th ACM International Conference on Information and Knowledge Management. Virtual Event, France, 2020, pp. 1–4. DOI: 10.1145/3340531.3417460. URL: https://hal.inria.fr/hal-03004218.

[13] A. Pacaci, A. Bonifati and T. M. Özsu. ‘Regular Path Query Evaluation on Streaming Graphs’. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. SIGMOD ’20. Portland, United States, June 2020, pp.1415–1430. DOI: 10.1145/3318464.3389733. URL: https://hal.archives-ouvertes.fr/hal-03118350.

Conferences without proceedings

[14] 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.

[15] A. Bonifati, U. Comignani and E. Tsamoura. ‘Exchanging Data under Policy Views’. In: EDBT 2021 - 24th International Conference on Extending Database Technology. Nicosia, Cyprus: https://edbticdt2021.cs.ucy.ac.cy/, 23rd Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03149043.

Scientific book chapters

[16] D. Schwab, S. Riou, A. Fejza, L. Vial, J. Marku, W. E. Husseini, E. K. Sannara, M. Bardon and Y. Robert. ‘Le projet GazePlay : des jeux ouverts, gratuits et une communauté pour les personnes en situation de polyhandicap’. In: *1024 – Bulletin de la Société informatique de France*. 1st Apr. 2020. URL: https://hal.archives-ouvertes.fr/hal-03004915.

Reports & preprints

[17] A. Bonifati, S. Dumbrava and H. Kondylakis. *Graph Summarization*. Apr. 2020. URL: https://hal.inria.fr/hal-03128573.

[18] 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.

[19] 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.

[20] O. Issa, A. Bonifati and F. Toumani. *A relational framework for inconsistency-aware query answering*. 9th Sept. 2020. URL: https://hal.archives-ouvertes.fr/hal-02934283.

[21] C. Koutras, G. Siachamis, A. Ionescu, K. Psarakis, J. Brons, M. Fragkoulis, C. Lofi, A. Bonifati and A. Katsifodimos. *Valentine: Evaluating Matching Techniques for Dataset Discovery*. Oct. 2020. URL: https://hal.inria.fr/hal-03128590.

[22] 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*. Nov. 2020. DOI: 10.1145/3434642. URL: https://hal.inria.fr/hal-03128601.

Other scientific publications

[23] L. S. Werner. ‘Predictive Analytics by Inferring Structure from Electronic Health Records’. Karlsruhe Institut für Technologie (KIT), 1st Sept. 2020. URL: https://hal.inria.fr/hal-03125018.