RESEARCH CENTRE

Lille - Nord Europe

IN PARTNERSHIP WITH:

CNRS, Université de Lille

2021 ACTIVITY REPORT

Project-Team LINKS

Linking Dynamic Data

IN COLLABORATION WITH: Centre de Recherche en Informatique, Signal et Automatique de Lille

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Contents

Project-Team LINKS

Creation of the Project-Team: 2016 June 01

Keywords

Computer sciences and digital sciences

A2.1. – Programming Languages

A2.1.1. – Semantics of programming languages

A2.1.4. – Functional programming

A2.1.6. – Concurrent programming

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

A2.4.1. – Analysis

A2.4.2. – Model-checking

A2.4.3. – Proofs

A3.1. – Data

A3.1.1. – Modeling, representation

A3.1.2. – Data management, quering and storage

A3.1.3. – Distributed data

A3.1.4. – Uncertain data

A3.1.5. – Control access, privacy

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

A3.2.2. – Knowledge extraction, cleaning

A3.2.3. – Inference

A3.2.4. – Semantic Web

A4.7. – Access control

A4.8. – Privacy-enhancing technologies

A7. – Theory of computation

A7.2. – Logic in Computer Science

A9.1. – Knowledge

A9.2. – Machine learning

A9.7. – AI algorithmics

A9.8. – Reasoning

Other research topics and application domains

B6.1. – Software industry

B6.3.1. – Web

B6.3.4. – Social Networks

B6.5. – Information systems

B9.5.1. – Computer science

B9.5.6. – Data science

B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientists

- Joachim Niehren [Team leader, Inria, Senior Researcher, HDR]
- Mikael Monet [Inria, Researcher]

Faculty Members

- Iovka Boneva [Université de Lille, Associate Professor]
- Florent Capelli [Université de Lille, Associate Professor]
- Aurélien Lemay [Université de Lille, Associate Professor, HDR]
- Charles Paperman [Université de Lille, Associate Professor]
- Sylvain Salvati [Université de Lille, Professor, HDR]
- Slawomir Staworko [Université de Lille, Associate Professor, HDR]
- Sophie Tison [Université de Lille, Professor, HDR]

PhD Students

- Antonio Al Serhali [Inria]
- Corentin Barloy [Université de Lille]
- Nicolas Crosetti [Inria]
- Paul Gallot [Université de Lille]
- Claire Soyez-Martin [Inria]

Technical Staff

- Cherif Amadou Ba [Inria, Engineer, until Jun 2021]
- Momar Ndiouga Sakho [Inria, Engineer, until Apr 2021]

Interns and Apprentices

- Zakaria Boulkhir [Université de Lille, from Apr 2021 until Jun 2021]
- Laurine Dargaud [École centrale de Lille, from Apr 2021 until Jun 2021]

Administrative Assistant

- Nathalie Bonte [Inria]

Visiting Scientist

- Corentin Barloy [École Normale Supérieure de Paris, from Oct 2020 until Oct 2021]

2 Overall objectives

We develop algorithms for answering logical querying on heterogeneous linked data collections in hybrid formats, distributed programming languages for managing dynamic linked data collections and workflows based on queries and mappings, and symbolic machine learning algorithms that can link datasets by inferring appropriate queries and mappings.

2.1 Presentation

The following three items summarize our main research objectives.

Querying Heterogeneous Linked Data We develop new kinds of schema mappings for semi-structured datasets in hybrid formats including graph databases, RDF collections, and relational databases. These induce recursive queries on linked data collections for which we will investigate evaluation algorithms, containment problems, and concrete applications.

Managing Dynamic Linked Data In order to manage dynamic linked data collections and workflows, we will develop distributed data-centric programming languages with streams and parallelism, based on novel algorithms for incremental query answering, study the propagation of updates of dynamic data through schema mappings, and investigate static analysis methods for linked data workflows.

Linking Data Graphs Finally, we will develop symbolic machine learning algorithms, for inferring queries and mappings between linked data collections in various graphs formats from annotated examples.

3 Research program

3.1 Background

The main objective of LINKS is to develop methods for querying and managing linked data collections. Even though open linked data is the most prominent example, we will focus on hybrid linked data collections, which are collections of semi-structured datasets in hybrid formats: graph-based, RDF, relational, and NOSQL. The elements of these datasets may be linked, either by pointers or by additional relations between the elements of the different datasets, for instance the "same-as" or "member-of" relations as in RDF.

The advantage of traditional data models is that there exist powerful querying methods and technologies that one might want to preserve. In particular, they come with powerful schemas that constraint the possible manners in which knowledge is represented to a finite number of patterns. The exhaustiveness of these patterns is essential for writing of queries that cover all possible cases. Pattern violations are excluded by schema validation. In contrast, RDF schema languages such as RDFS can only enrich the relations of a dataset by new relations, which also helps for query writing, but which cannot constraint the number of possible patterns, so that they do not come with any reasonable notion of schema validation.

The main weakness of traditional formats, however, is that they do not scale to large data collections as stored on the Web, while the RDF data models scales well to very big collections such as linked open data. Therefore, our objective is to study mixed data collections, some of which may be in RDF format, in which we can lift the advantages of smaller datasets in traditional formats to much larger linked data collections. Such data collections are typically distributed over the internet, where data sources may have rigid query facilities that cannot be easily adapted or extended.

The main assumption that we impose in order to enable the logical approach, is that the given linked data collection must be correct in most dimensions. This means that all datasets are well-formed with respect to their available constraints and schemas, and clean with respect to the data values in most of the components of the relations in the datasets. One of the challenges is to integrate good quality RDF datasets into this setting, another is to clean the incorrect data in those dimensions that are less proper. It remains to be investigated in how far these assumptions can be maintained in realistic applications, and how much they can be weakened otherwise.

For querying linked data collections, the main problems are to resolve the heterogeneity of data formats and schemas, to understand the efficiency and expressiveness of recursive queries, that can follow links repeatedly, to answer queries under constraints, and to optimize query answering algorithms based on static analysis. When linked data is dynamically created, exchanged, or updated, the problems are how to process linked data incrementally, and how to manage linked data collections that change dynamically. In any case (static and dynamic) one needs to find appropriate schema mappings for linking semi-structured datasets. We will study how to automatize parts of this search process by developing symbolic machine learning techniques for linked data collections.

3.2 Research Axis: Querying Data Graphs

Linked data is often abstracted as datagraphs: nodes carry information and edges are labeled. Internet, the semantic web, open data, social networks and their connections, information streams such as twitter are examples of such datagraphs. An axis of LINKS is to propose methods and tools so as to extract information from datagraphs. We dwell in a wide spectrum of tools to construct these methods: circuits, compilation, optimization, logic, automata, machine learning. Our goal is to extend the kinds of information that can be extracted from datagraphs while improving the efficiency of existing ones.

This axis is split within two themes. The first one pertains to the use of low level representations by means of circuits to compute efficiently complex numerical aggregates that will find natural applications in AI. The second one proposes to explore path oriented query language and more particularly their efficient evaluation by means of efficient compilation and machine learning methods so as to have manageable statistics.

3.2.1 AI: Circuits for Data Analysis

Circuits are concise representations of data sets that recently found a unifying interest in various areas of artificial intelligence. A circuit may for instance represent the answer set of a database query as a dag whose operators are disjoint unions (for disjunction) and Cartesian products (for conjunction). Similarly, it may also represent the set of all matches of a pattern in a graph. The structure of the circuit may give rise to efficient algorithms to process large data sets based on representation that are often much smaller. Among others, such applications range from knowledge representation/compilation, counting the number of solutions of queries, efficient query answering, factorized databases.

In a first line of research, we want to study novel problems on circuits, in which database queries are relevant to data analysis tasks from artificial intelligence, in machine learning or data mining in particular. In particular we propose to study optimization problems on answer sets of database queries based on circuits, i.e. how to find optimal solutions in the answer set for a given set of conditions. Decompressing small circuits into large answer sets would make the optimization problem unfeasible in many cases. We believe that circuits can structure certain optimization problems in such a way that it can be phrased concisely and then solved efficiently.

Second, we propose to develop a tighter integration between circuits and databases. Indeed query-related circuits are generally produced from a database. This requires that the data is copied within the circuits. This memory cost is accompanied with the loss of the environment of the DBMS which allows many optimizations and uses many low level optimizations that are hard to implement. We propose then to encode circuits directly within the database using materialized views and index structures. We shall also develop the required computational tools for maintaining and exploiting these embedded circuits.

3.2.2 Path Query Optimization

Graph databases are easily queried using path descriptions. Most often these paths are described by means of regular expressions. This makes path queries difficult as the use of Kleene star makes them recursive. In relational DBMS, recursion is almost never used and it is not advised to use it. The natural theoretical tool that pertains to recursion in the context of relational data Datalog. There has been a wealth of optimization algorithms that have been proposed for queries written in Datalog. We propose to use Datalog as a low level language to which we will compile path queries of various kinds. The idea is that the compiler will try to obtain Datalog programs that will have low execution complexity

taking advantages of optimization techniques such as magic supplementary set rewriting, pre-computed indexes and also statistics computed from the graph. Our goal is to develop a compiler that will be able to efficiently evaluate path queries on large graphs which in particular will explore only a part of it.

3.3 Research Axis: Monitoring Data Graphs

Traditional database applications are programs that interact with database via updates and queries. We are interested in developing programming language techniques so as to interact with datagraphs rather than with traditional relational databases. Moreover, we shall take into account the dynamic aspects of datagraphs which shall evolve through updates. We will develop methods to monitor changes in datagraphs and react according to the modifications.

3.3.1 Functional Programming Languges for Data Graphs

The first question is which kind of programming language to use to enable monitoring processes for data graphs based on query answering. While languages of path queries found quite some interest on data graphs, less attention has been given to the programming language tasks, that needed to be solved to produce structured output and to compose various queries with structured output into a pipeline. We believe that transferring the generalization of ideas developed for data trees in the context of XML to data graphs will allow to solve such problems in a systematic manner.

Our approach will consist in developing a functional programming language based on first principles (the lambda calculus, graph navigation, logical connective) that generalizes full XPath 3.0 to the context of graphs. Here we can rely on own previous work for data trees, such as the language X-Fun and λ-XP. After the language for data graphs is designed we shall study its behavior empirically by means of an implementation. This implementation will help us to design optimization methods so as to evaluate the queries in that language. This will allow us to use a wealth of techniques so as to optimize the computation. Indeed, we can try to compile data structures to imperative ones when possible and also exploit possibilities of parallel executions in certain cases. Functional programming comes with nice verification techniques that we are going to use in several contexts: (i) in optimizing queries (e.g. stop the evaluation when it is possible to know that no more data can contribute to the output) and (ii) to verify that the query behaves correctly. The verification methods we shall focus on will be mainly related to automata and transducers.

Finally we shall also develop a programming language that allows to describe services that use datagraphs as a backend for storing data. Here again, functional programming seems a good candidate, we would need however to orchestrate the concurrent executions of queries so as to ensure the correct behavior of services. This means that we should have concurrent constructs that are built in the language. The high level of concurrence enabled by the notion of *futures* seems an interesting candidate to adapt to the context of service orchestration.

3.3.2 Hyperstreaming Program Evaluation

Complex-event processing requires to monitor data graphs that are produced on input streams and to write data graphs to some output stream, which can then be used as inputs again. A major problem here is to reduce the high risk of blocking, which arises when the writing of some of the output stream suspends on a data value that will become available only in the future on some input stream. In such cases, all monitoring processes reading the output stream may have to suspend as well. In order to reduce the risk of blocking, we propose to develop the hyperstreaming approach further, of which we layed the foundations in the evaluation period based on automata techniques. The idea is to generalize streams to hyperstreams, i.e. to add holes to streams that can be filled by some other stream in the future. In order to avoid suspension as possible, a monitoring process on hyperstream must then be able to jump over the holes, and to perform some speculative computation. The objective for the next period are to develop tools for hyperstreaming query answering and to lift these to hyperstreaming program evaluation. Furthermore, on the conceptual side, the notion of certain query answers on hyperstreams needs to be lifted to certain program outputs on hyperstreams.

3.4 Research Axis: Graph Data Integration

We intend to continue to develop tools for integration of linked data with RDF being their principal format. Because from its conception the main credo of RDF has been "just publish your data", the problem at hand faces two important challenges: data quality and data heterogeneity.

3.4.1 Data Quality with Schemas and Repairing with Inference

The data quality of RDF may suffer due to a number of reasons. Impurities may arise due to data value errors (misspellings, errors during data entry etc.). Such data quality problems have been thoroughly investigated in literature for relational databases and solutions include dictionary methods etc. However, it remains to be seen if the challenges of adapting the existing solutions for relational databases can be easily addressed.

One particular challenge comes from the fact that RDF allows a higher degree of structural freedom in how information is represented as opposed to relation databases, where the choice is strongly limited to flat tables. We plan to investigate suitability of existing data cleaning methods to tackle the problems of data value impurities in RDF. The structural freedom of RDF is a source of data quality issues on its own. With the recent emergence of schema formalisms for RDF, it becomes evident that significant parts of existing RDF repositories do not necessarily satisfy schemas prepared by domain experts.

In the first place, we intend to investigate defining suitable measures of quality for RDF documents. Our approaches will be based on a schema language, such as ShEx and SHACL, and we shall explore suitable variants of graph alignment and graph edit distance to capture similarity between the existing RDF document and its possible repaired versions that satisfy the schema.

The central issue here is repairing an RDF document w.r.t. schema by identifying essential fragments of the RDF that fail to satisfy the schema. Once such fragments are identified, repairing actions can be applied however there might be a significant number of alternatives. We intend to explore enumeration approaches where the space of repairing alternatives is intelligently browsed by the user and the most suitable one is chosen. Furthermore, we intend to propose a rule language for choosing the most suitable repairing action and will investigate inference methods to derive from interactions with user the optimal order in which various repairing actions are presented to the user and derive the rules for the choice of the preferred repairing action for repeating types of fragments that do not satisfy the schema.

3.4.2 Integration and Graph Mappings with Schemas and Inference

The second problem pertaining to integration of RDF data sources is their heterogeneity. We intend to continue to identify and study suitable classes of mappings between RDF documents conforming to potentially different and complementary schemas. We intend to assist the user in constructing such mappings by developing rich and expressive graphical languages for mappings. Also, we wish to investigate inference of RDF mappings with the active help of an expert user. We will need to define interactive protocols that allows the input to be sufficiently informative to guide the inference process while avoiding the pitfalls of user input being too ambiguous and causing combinatorial explosion. We intend to identify

RDF Data Quality. Approach based on a schema language (ShEx or SHACL) used to identify errors and giving a notion of a measure of quality of an RDF database. Impurities in RDF may come from data value errors (misspellings etc.) but also from the fact that RDF imposes fewer constraints on how data is structured which is a consequence of a significantly different use philosophy (just publish your data anyway you want). Repairing of RDF errors would be modeled with a localized rules (transformations that operate within a small radius of an affected node) and if several rules apply, preferences are used to identify the most desirable one. Both the repairing rules and preferences can be inferred with the help of inference algorithms in an interactive setting. Smart tools for LOD integration. Assuming that the LOD sources are of good quality, we want to build tools that assist the user in constructing mappings that integrate data in the user database. For this, we want to define inference algorithms which are guided by schemas, and which are based on comprehensible interactions with the user. For this, we need to define interactions that are rich enough to inform the algorithm, while simple enough to be understandable by a non-expert user. In particular, that means that we need to present data (nodes in a graph for instance)

in a readable way. Also, we want to investigate how the - possibly inferred - schema can be used to guide the inference.

4 Application domains

4.1 Linked data integration

There are many contexts in which integrating linked data is interesting. We advocate here one possible scenario, namely that of integrating business linked data to feed what is called Business Intelligence. The latter consists of a set of theories and methodologies that transform raw data into meaningful and useful information for business purposes (from Wikipedia). In the past decade, most of the enterprise data was proprietary, thus residing within the enterprise repository, along with the knowledge derived from that data. Today's enterprises and businessmen need to face the problem of information explosion, due to the Internet's ability to rapidly convey large amounts of information throughout the world via end-user applications and tools. Although linked data collections exist by bridging the gap between enterprise data and external resources, they are not sufficient to support the various tasks of Business Intelligence. To make a concrete example, concepts in an enterprise repository need to be matched with concepts in Wikipedia and this can be done via pointers or equalities. However, more complex logical statements (i.e. mappings) need to be conceived to map a portion of a local database to a portion of an RDF graph, such as a subgraph in Wikipedia or in a social network, e.g. LinkedIn. Such mappings would then enrich the amount of knowledge shared within the enterprise and let more complex queries be evaluated. As an example, businessmen with the aid of business intelligence tools need to make complex sentimental analysis on the potential clients and for such a reason, such tools must be able to pose complex queries, that exploit the previous logical mappings to guide their analysis. Moreover, the external resources may be rapidly evolving thus leading to revisit the current state of business intelligence within the enterprise.

4.2 Data cleaning

The second example of application of our proposal concerns scientists who want to quickly inspect relevant literature and datasets. In such a case, local knowledge that comes from a local repository of publications belonging to a research institute (e.g. HAL) need to be integrated with other Web-based repositories, such as DBLP, Google Scholar, ResearchGate and even Wikipedia. Indeed, the local repository may be incomplete or contain semantic ambiguities, such as mistaken or missing conference venues, mistaken long names for the publication venues and journals, missing explanation of research keywords, and opaque keywords. We envision a publication management system that exploits both links between database elements, namely pointers to external resources and logical links. The latter can be complex relationships between local portions of data and remote resources, encoded as schema mappings. There are different tasks that such a scenario could entail such as (i) cleaning the errors with links to correct data e.g. via mappings from HAL to DBLP for the publications errors, and via mappings from HAL to Wikipedia for opaque keywords, (ii) thoroughly enrich the list of publications of a given research institute, and (iii) support complex queries on the corrected data combined with logical mappings.

4.3 Real-time complex event processing

Complex event processing serves for monitoring nested word streams in real time. Complex event streams are gaining popularity with social networks such as with Facebook and Twitter, and thus should be supported by distributed databases on the Web. Since this is not yet the case, there remains much space for future industrial transfer related to LINKS' second axis on dynamic linked data.

5 Social and environmental responsibility

5.1 Footprint of research activities

Tison Member of the general assembly of the European Association for Theoretical Computer Science (EATCS) (elected in 2019).

5.2 Impact of research results

Databases and methods from Artificial Intelligence are used in virtually all aspects of the modern digitalized world, from companies' web services to governments' institutions.

6 Highlights of the year

6.1 Linear Programs with Database Queries

Florent Capelli, Nicolas Crosetti, Joachim Niehren and Jan Ramon from Inria Magnet presented a novel language for defining linear programs based on database queries. They also found an efficient algorithm that can compute the optimal value of linear programs with database queries without materializing the queries' answer sets, while assuming queries of small hypertree width. This optimization algorithm can be applied to minimizing noise in ε-differential privacy, and to computing the s-measure in data mining. These results are generally relevant to the field of artificial intelligence (AI). They will be presented at the International Conference on Database Theory 2022 [6] where they received the best newcomer award.

6.2 Spotting Bugs in Debian Installation Scripts

Paul Gallot [28] developed a novel algorithm based on tree transducers, for spotting bugs in Linux installation scripts. On a corpus of Debian package scripts, his implementation could find around 40 genuine bugs that were reported to the Debian community. These results are part of Gallot's PhD thesis directed by Sylvain Salvati and Aurélien Lemay, and was funded by the ANR project CoLiS coordinated by Ralf Treinen on the *Correctness of Linux Scripts*. The ColiS approach is to consider file systems as data trees, so that tree transducers can be used to represent operations on file system performed by Shell scripts to install and remove packages on Debian GNU/Linux distributions.

6.3 ICALP'2021 Best Paper Award

Paperman et al. received a best paper award for their ICALP'2021 article "Dynamic Membership for Regular Languages" [1], in which they study the *dynamic membership problem* for regular languages. This fundamental problem can be defined as follow: fix a language L, read a word w, build in time $O(|w|)$ a data structure indicating if w is in L, and maintain this structure efficiently under letter substitutions on w. Intuitively, one can see the flux of letter substitutions as a stream of information, and one wants to be able to efficiently determine if the word is still in the language or not. They obtain a trichotomy on the complexity of the problem – the description of which would be too technical for this document –, and solve related open questions on the dynamic word problem.

7 New software and platforms

7.1 New software

7.1.1 ShEx validator

Name: Validation of Shape Expression schemas

Keywords: Data management, RDF

Functional Description: Shape Expression schemas is a formalism for defining constraints on RDF graphs. This software allows to check whether a graph satisfies a Shape Expressions schema.

Release Contributions: ShExJava now uses the Commons RDF API and so support RDF4J, Jena, JSON-LD-Java, OWL API and Apache Clerezza. It can parse ShEx schema in the ShEcC, ShEJ, ShExR formats and can serialize a schema in ShExJ.

To validate data against a ShExSchema using ShExJava, you have two different algorithms: - the refine algorithm: compute once and for all the typing for the whole graph - the recursive algorithm: compute only the typing required to answer a validate(node,ShapeLabel) call and forget the results.

URL: http://shexjava.lille.inria.fr/

Contact: Iovka Boneva

7.1.2 gMark

Name: gMark: schema-driven graph and query generation

Keywords: Semantic Web, Data base

Functional Description: gMark allow the generation of graph databases and an associated set of query from a schema of the graph.gMark is based on the following principles: - great flexibility in the schema definition - ability to generate big size graphs - ability to generate recursive queries - ability to generate queries with a desired selectivity

URL: https://github.com/graphMark/gmark

Contact: Aurélien Lemay

7.1.3 SmartHal

Keyword: Bibliography

Functional Description: SmartHal is a better tool for querying the HAL bibliography database, while is based on Haltool queries. The idea is that a Haltool query returns an XML document that can be queried further. In order to do so, SmartHal provides a new query language. Its queries are conjunctions of Haltool queries (for a list of laboratories or authors) with expressive Boolean queries by which answers of Haltool queries can be refined. These Boolean refinement queries are automatically translated to XQuery and executed by Saxon. A java application for extraction from the command line is available. On top of this, we have build a tool for producing the citation lists for the evaluation report of the LIFL, which can be easily adapter to other Labs.

URL: http://smarthal.lille.inria.fr/

Contact: Joachim Niehren

7.1.4 QuiXPath

Keywords: XML, NoSQL, Data stream

Scientific Description: The QuiXPath tools supports a very large fragment of XPath 3.0. The QuiXPath library provides a compiler from QuiXPath to FXP, which is a library for querying XML streams with a fragment of temporal logic.

Functional Description: QuiXPath is a streaming implementation of XPath 3.0. It can query large XML files without loading the entire file in main memory, while selecting nodes as early as possible.

URL: https://project.inria.fr/quix-tool-suite/

Contact: Joachim Niehren

7.1.5 X-FUN

Keywords: Programming language, Compilers, Functional programming, Transformation, XML

Functional Description: X-FUN is a core language for implementing various XML, standards in a uniform manner. X-Fun is a higher-order functional programming language for transforming data trees based on node selection queries.

Contact: Joachim Niehren

Participants: Joachim Niehren, Pavel Labath

7.1.6 ShapeDesigner

Name: ShapeDesigner

Keywords: Validation, Data Exploration, Verification

Functional Description: ShapeDesigner allows construct a ShEx or SHACL schema for an existing dataset. It combines algorithms to analyse the data and automatically extract shape constraints, and to edit and validate shape schemas.

URL: https://gitlab.inria.fr/jdusart/shexjapp

Contact: Jeremie Dusart

7.1.7 Coussinet

Name: Coussinet

Keywords: Enumeration, Complexity

Functional Description: Coussinet is a demo illustrating a technique called geometric amortization for enumeration algorithms introduced in the paper Geometric Amortization for Enumeration Algorithms, Florent Capelli, Yann Strozecki. The result presented in this paper is about making the delay of enumeration algorithms more regular.

URL: http://florent.capelli.me/coussinet/coussinet.html

Contact: Florent Capelli

Participants: Florent Capelli, Yann Strozecki

7.1.8 Colis

Keywords: Debug, Automated deduction

Functional Description: Paul Gallot developped a novel algorithm based on tree transducers, for spotting bugs in linux installation scripts. On a corpus of Debian package scripts, his implementation could find around 40 genuine bugs that were reported to the Debian community. This software is is part of Paul's PhD thesis, directed by Sylvain Salvati and Aurélien Lemay, and was funded by the ANR projec CoLiS coordinated by Ralf Treinen on the Correctness of Linux Scripts. The ColiS approach is to consider file systems as data trees, so that tree transducers can be used to represent operations on file system performed by Shell scripts to install and remove packages on Debian GNU/Linux distributions.

Contact: Paul Gallot

8 New results

8.1 Querying Data Graphs

8.1.1 Circuits for Data Analysis in Artificial Intelligence

Knowledge compilation is a general technique in artificial intelligence to obtain tractable algorithms for subclasses of algorithmic problems that are computationally hard. For instance, a variant of Yannakakis' algorithm can be used to compile acyclic conjunctive database queries to Boolean circuits. These will then be decomposable and deterministic, and thus tractable in polynomial time, while for the general class of conjunctive queries, testing the existence of a query answer on a relational database is coNP-complete. Another class of instances, where knowledge complilation is used in AI, concern satisfiablity problems. Beside of satisfiability, knowledge compilation is equally relevant to aggregation and enumeration problems.

Capelli et al. present at AAAI [25] a method to certify the output knowledge compilers and #SAT-solvers. This is a cooperation with Université de Lens. The idea is to output a certificate that can be checked in polynomial time and can be used to certify that a given CNF formula has K models. Their experiments were encouraging showing that a large majority of CNF formulas for which the #SAT-solver D4 terminates have certificates that can be checked more quickly than the compilation time.

In their article recently accepted at ICDT'2022 [24], Capelli, Crosetti, Niehren and Ramon study the problem of optimizing a linear program whose variables are answers to a conjunctive query. For this they propose a new language for specifying linear programs whose constraints and objective functions depend on the answer sets of conjunctive queries. They developed an efficient algorithm for solving programs in a fragment of this language. Using tools from knowledge compilation, and exploiting the structure of queries having small *treewidth* (a graph parameter, intuitively measuring how far a graph is from being a tree), they are able to constructs a linear program having the same optimal value but fewer variables, thus nontrivially improving the asymptotic complexity of solving this task. They moreover illustrate the application of their language on three examples: optimizing deliveries of resources, minimizing noise for differential privacy, and computing the s-measure of patterns in graphs as needed for data mining.

8.1.2 Uncertainty and Explanations

Monet et al. [22] study in a paper at AAAI Shapely values for providing explanations to classification results over machine learning models. This work is done in cooperation with the Pontifical Catholic University of Chile. While in general computing Shapley values is a computationally intractable problem, it has recently been claimed that the SHAP-score can be computed in polynomial time over the class of decision trees. They show that the SHAP-score can be computed in polynomial time over deterministic and decomposable Boolean circuits.

In a new article in Transactions on Computational Logic [13], Monet et al. study the complexity of various counting problems that arise in the context of incomplete databases, i.e., relational databases that can contain unknown values in the form of labeled nulls. Intuitively, these problems can be used to answer queries on such incomplete data, for instance by computing how many ways there are to associate constants to the missing values of the database such that a given query is satisfied. They prove complexity dichotomy results between #P-hardness (i.e., highly intractable) and polynomial-time computability for these problems for self-join–free conjunctive queries (intuitively corresponding to the SELECT-PROJECT-JOIN fragment of SQL) and study the impact on the complexity of the various natural restrictions. They also study the *approximability* of these problems, and consider more expressive query languages to situate these problems with respect to known complexity classes.

In a recently accepted paper at SIGMOD'2022 [26], Monet et al. use the framework of *Shapley values* to assign and compute contributions of input facts of a database to the results of a query. The goal is, intuitively, to *explain* the results of a query by computing a score for every input fact. The Shapley value is a game-theoretic notion for wealth distribution that is nowadays extensively used to explain complex data-intensive computation, for instance, in network analysis or machine learning. Monet et al. present

in this paper two practically effective solutions for computing Shapley values in query answering. First, they establish a tight theoretical connection to the extensively studied problem of *query evaluation over probabilistic databases*, which allows then to obtain a polynomial-time algorithm for the class of queries for which probability computation is tractable. They then propose a first practical solution for computing Shapley values that adopts tools from probabilistic query evaluation and knowledge compilation. Experiments are caried that demonstrate the practical effectiveness of their solutions.

8.1.3 Query Optimization

In a PVLDB article [15], Staworko et al. study *threshold queries*, that is, queries that only require computing or counting answers up to a specified threshold value. This type of query is very common in practice, yet has been little studied. In this paper, they present a theoretical analysis of threshold query evaluation and show that thresholds can be used to significantly improve the asymptotic bounds of state-of-the-art query evaluation algorithms. In surprising contrast to conventional wisdom, they found important scenarios in real-world data sets in which users are interested in computing the results of queries up to a certain threshold, independent of a ranking function that orders the query results by importance.

8.2 Monitoring Data Graphs

8.2.1 Functional Programming Languages for Data Trees

Paul Gallot defended his thesis, “Safety of transformations of data trees: Tree transducer theory applied to a verification problem on shell scripts” [28], on December 16th 2021. Part of this thesis presents work done in the scope of the CoLiS project (ANR), which stands for *Correctness of Linux Scripts*. CoLiS’s stated goal is to apply techniques from deductive program verification and analysis of tree transformations to the problem of analyzing Shell scripts used in software installation. Paul’s thesis aims at studying formal modelisations of tree transformations, with a focus on tree transducers. In particular he uses tree transformations to represent operations on file systems which are represented as tree structures, by modelising operations performed by Shell scripts used to install and remove packages on Debian GNU/Linux distributions. Unix file systems are modeled as feature trees and the actions of Unix commands on a file system can be represented using a model he calls tree pattern transducers. He proposes to translate Unix commands into this model, and then then provides an algorithm for computing the composition of tree pattern transducers. A tool for finding the configurations of the file system which can make a given Shell script fail is implemented. The implementation is then tested on a corpus of Debian package scripts, and around 40 genuine bugs were found and reported to the Debian community. The thesis also contains more theoretical results, for instance the use of techniques from the field of functional programming to shed new light on known models of transducers.

8.2.2 Query Answering on Streams

In an article published at Algorithms [18] extending on a paper published at CSR’2020, they could show that regular path queries on XML documents in the usual XPathMark benchmark can be compiled to reasonably small deterministic automata on nested words. For this they propose new compilers to the novel class of deterministic stepwise hedge automata and proposed a minimization algorithm for them. We note that streaming evaluators for such automata are heavily stack based.

Paperman et al. received a best paper award for their ICALP’2021 article “Dynamic Membership for Regular Languages” [20], in which they study the *dynamic membership problem* for regular languages. This fundamental problem can be defined as follow: fix a language L, read a word w, build in time $O(|w|)$ a data structure indicating if w is in L, and maintain this structure efficiently under letter substitutions on w. Intuitively, one can see the flux of letter substitutions as a stream of information, and we want to be able to efficiently determine if the word is still in the language or not. They obtain a trichotomy on the complexity of the problem – the description of which would be too technical for this document –, and

solve related open questions on the dynamic word problem.

Boneva, Niehren et al. [14] study the complexity of regular matching and inclusion for compressed tree patterns with context variables subject to regular constraints. Context variables with regular constraints permit to properly generalize on unranked tree patterns with hedge variables. Regular inclusion on unranked tree patterns is relevant to certain query answering on Xml streams with references. They prove that regular matching and inclusion with regular constraints can be reduced in polynomial time to the corresponding problem without regular constraints.

8.3 Graph Data Integration

8.3.1 Data Quality with Schemas and Repairing with Inference

Slawomir Staworko is part of a community effort between industry and academia to shape the future of property graph constraints, the results of which are reported in a recent SIGMOD paper [21]. The standardization for a property graph query language is currently underway through the ISO Graph Query Language (GQL) project. Their position is that this project should pay close attention to schemas and constraints, and especially on key constraints. Motivated by use cases from the industry partners, they argue that key constraints should be able to have different modes, which are combinations of basic restriction that require the key to be exclusive, mandatory, and singleton. This lead them to propose *PG-Keys*, a flexible and powerful framework for defining key constraints that aims to guide the evolution of the standardization efforts towards making systems more useful, powerful, and expressive.

8.3.2 Integration and Graph Mappings with Schemas and Inference

In an ICDT'2022 paper [27], Lemay and Staworko, together with Benoît Groz and Piotr Wieczorek, investigate the problem of constructing a shape graph that describes the structure of a given graph database. They employ the framework of grammatical inference, where the objective is to find an inference algorithm that is both sound, i.e., always producing a schema that validates the input graph, and complete, i.e., able to produce any schema, within a given class of schemas, provided that a sufficiently informative input graph is presented. They identify a number of fundamental limitations that preclude feasible inference, and present inference algorithms based on natural approaches that allow to infer schemas that they argue to be of practical importance.

8.4 Others

In a Theory of Computing Systems article [16], Paperman et al. study the expressive power of polynomial recursive sequences, which are nonlinear extensions of the well-known class of linear recursive sequences. These sequences arise naturally in the study of nonlinear extensions of weighted automata, where (non)expressiveness results translate to class separations. A typical example of a polynomial recursive sequence is $b_n = n!$. The main result is that the sequence $u_n = n^n$ is not polynomial recursive.

Niehren is cooperating with the BioComputing team of the Cristal Lab at the University of Lille since many years. He uses abstract interpretation of logical formulas for predicting gene knockouts based on formal models of reaction network. Applications of Niehren's prediction algorithms from systems biology for surfactine overproduction are discussed in a survey of the European project BestBioSurf [19]. In cooperation with bio-engineers from Clermont Ferrant and Lille, Niehren presented a novel application of his prediciton algorithms to the overproduction of Mycosubtilin isoforms [17]. This required an extension to the prediction of gene-overexpressions. Niehren improved the precision of the general prediction algorithm in [11] with his colleagues from BioComputing. For this, they showed how to compute the difference abstraction of linear equation systems exactely. Previous algorithms could only compute over-approximations. Exactness yields a completeness result for abstract interpretation, while over-approximations justify the correctness. The exact computation of difference abstractions is based on the exact rewriting of linear equations systems with respect to the boolean abstraction [12]. The latter can be computed from the elementary modes of the linear equation system.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

Stawarko Academic member of Linked Data Benchmark Council (LDBC).

Staworko Member of Work Group on Property Graph Schemas (standardisation effort).

Tison Vice-president of the Force Awards association.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Participation in other International Programs

Informal International Partners

Santiago, Chile Monet cooperates with Marcelo Arenas and Pablo Berceló from Pontificia Universidad Católica de Chile and with Luca Bertossi from Universidad Adolfo Ibanez (also Chile) on counting problems for incomplete databases and on the computation of SHAP-score explainations for circuit classes from knowledge compilation. This yield a joint publication at ACM Transactions on Computational Logic [13].

Tel Aviv, Israel Monet works with Benny Kimelfeld from Technion (Israel) and Daniel Deutch from Tel Aviv University on computing Shapley values for database query answers. A joint paper has been accepted at at ACM SIGMOD/PODS 2022 [26].

Warsaw, Poland Paperman cooperates with Filip Murlak on query evaluation on streams. A joint paper was accepted for publication at PODS'2021 [23].

Wroclaw, Poland Staworko has regular exchange with Piotr Wieczorek from the University of Wroclaw, which lead to a joined publication at PODS 2019.

Saint Petersburg, Russia Salvati and Niehren cooperate with the University of Saint Petersburg following a visit of R. Azimov. This cooperation was funded by a invitation for R. Azimov by the Cristal lab in 2019.

10.2 International research visitors

10.2.1 Visits of international scientists

Nofar Cameli Technion, Israel. Links' online seminar. Nov, 2021.

Sebastian Maneth Bremen University, Germany. Dec 16, 2021.

10.3 National initiatives

ANR JCJC KCODA

Participants: Florent Capelli *(correspondent)*, Charles Paperman, Sylvain Salvati.

- **Duration:** 2021–2025
- **Objectives:** The aim of KCODA is to study how succinct representations can be used to efficiently solve modern optimization and AI problems that use a lot of data. We suggest using data structures from the field of compilation of knowledge that can represent large datasets succinctly by factoring certain parts while allowing efficient analysis of the represented data. The first goal of KCODA is to

understand how one can efficiently solve optimization and training problems for data represented by these structures. The second goal of KCODA is to offer better integration of these techniques into the systems of database management by proposing new algorithms allowing to build factorized representations of the data responses to DB requests and by proposing encodings of these representations inside the DB.

ANR Colis — Correctness of Linux Scripts

Participants: Joachim Niehren *(correspondent)*, Aurélien Lemay, Paul Gallot, Sylvain Salvati.

- **Duration:** 2015–2021
- **Coordinator:** R. Treinen, Université Paris Diderot
- **Partner:** C. Marché, Tocata project-team, Inria Saclay.
- **Objective:** This project aims at verifying the correctness of transformations on data trees defined by shell scripts for Linux software installation. The data trees here are the instance of the file system which are changed by installation scripts.

ANR DataCert

Participants: Iovka Boneva *(correspondent)*, Sophie Tison, Jose Martin Lozano.

- **Duration:** 2015–2021
- **Coordinator:** E. Contejean, Université Paris-Sud
- **Partners:** Université de Lyon
- **Objective:** The main goals of the DataCert project are to provide deep specification in Coq of algorithms for data integration and exchange and of algorithms for enforcing security policies, as well as to design data integration methods for data models beyond the relational data model.

ANR Headwork

Participants: Joachim Niehren *(correspondent)*, Momar Ndiouga Sakho, Nicolas Crosetti, Florent Capelli.

- **Duration:** 2016–2022
- **Coordinator:** D. Gross-Amblard, Druid Team, Université de Rennes 1
- **Scientific partners:** Dahu project-team (Inria Saclay) and Sumo project-team (Inria Bretagne)
- **Industrial partners:** Spipoll and Foulefactory.
- **Objective:** The main object is to develop data-centric workflows for programming crowd-sourcing systems in flexible declarative manner. The problem of crowd-sourcing systems is to fill a database with knowledge gathered by thousands or more human participants. A particular focus is to be put on the aspects of data uncertainty and for the representation of user expertise.

ANR Delta

Participants: Joachim Niehren *(correspondent)*, Sylvain Salvati, Aurélien Lemay.

- **Duration:** 2016–2021
- **Partners:** LIF (Université Aix-Marseille) and IRIF (Université Paris-Diderot)
- **Coordinator:** M. Zeitoun, LaBRI (Université de Bordeaux)
- **Objective:** Delta is focused on the study of logic, transducers and automata. In particular, it aims at extending classical framework to handle input/output, quantities and data.

ANR Bravas

Participants: Sylvain Salvati *(correspondent)*.

- **Duration:** 2017–2022
- **Coordinator:** Jérôme Leroux, LaBRI, Université de Bordeaux
- **Scientific Partner:** LSV, ENS Cachan
- **Objective:** The goal of the BraVAS project is to develop a new and powerful approach to decide the reachability problems for Vector Addition Systems (VAS) extensions and to analyze their complexity. The ambition here is to crack with a single hammer (ideals over well-orders) several long-lasting open problems that have all been identified as a barrier in different areas, but that are in fact closely related when seen as reachability.

10.4 Regional initiatives

CPER Cornelia on Artificial Intelligence (2021-2025)

Participants: Joachim Niehren *(correspondent)*
.

The whole Links' project is partner of this new CPER project.

PhD project Nicolas Crosetti (2018-...) Cofunded by the Region Haut de France. In coopertion with Jan Ramon from Inria Magnet.

Participants: Sophie Tison *(supervisor)*, Florent Capelli, Joachim Niehren.

PhD project Antonio al Serhali (2020-...) Cofunded by the Region Haut de France.

Participants: Joachim Niehren.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of organizing committees

Niehren Organisation of the visit of students of ENS Paris-Sacaly at Cristal and Inria Lille, December 2021.

11.1.2 Scientific events: selection

Member of conference program committees

Boneva Member of the International Symposium on String Processing and Information Retrieval (SPIRE'2021) program committee.

Capelli Member of the AAAI Conference on Artificial Intelligence (AAAI'2021) program committee.

Capelli Member of the International Joint Conference on Artificial Intelligence (IJCAI-ECAI'2021) program committee.

Lemay Member of the International Conference on Grammatical Inference (ICGI'2022) program committee.

Monet Member of the AAAI Conference on Artificial Intelligence (AAAI'2021) program committee.

Monet Member of the ACM SIGMOD/PODS'2022 program committee.

Monet Member of the International Conference on Database Theory (ICDT'2022) program committee.

Staworko Member of the ACM SIGMOD/PODS'2022 program committee.

11.1.3 Journal

Member of editorial boards

Niehren Editorial Board of Fundamenta Informaticae.

Niehren Editorial Board of Algorithms.

Salvati Managing Editor of the Journal of Logic, Language and Information (JLLI).

Tison Editorial Board of RAIRO-ITA (until October 2021).

11.1.4 Scientific expertise

Capelli Member of Inria Lille CER (Commission des Emplois de Recherche)

Salvati Member of Inria's Evaluation Committee.

Tison Elected member of CNU 27.

Tison Expertise for the Austrian Science Fund (FWF).

11.1.5 Research administration

Salvati Member of the joint and restricted commissions of the computer science department of Université de Lille (FIL) and of CRIStAL for recruitments.

Staworko Member of the Parity/Equality commission of the CRIStAL laboratory.

Tison Member of the coordinating team of ISite Université de Lille - Nord Europe.

Tison Elected member of the executive board of Université de Lille.

Tison Member of the school board of IMT Nord Europe.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching Responsibilities

Capelli Responsible for the L1, LEA department, Université de Lille.

Capelli Elected member of the board of LEA department, Université de Lille.

Capelli Reponsible for Parcoursup for LEA department, Université de Lille.

Salvati Co-director of studies for the Master MIAGE FA, Université de Lille.

Salvati Director of studies for the mathematics and computer science bachelor's degree of Université de Lille.

Salvati Co-responsible for the research track of the computer science bachelor's degree of Université de Lille.

Salvati Board member of the computer science departement of Université de Lille (FIL).

Staworko Coordinator of International Relationships at the Department of Computer Science, Université de Lille (FIL).

Tison Member of the selection board for «Capes» in computer science.

11.2.2 Teaching Activities

Boneva Teaches computer science in DUT Informatique of Université de Lille

Capelli Teaches computer science in the LEA department of Université de Lille for around 200h per year (Licence and Master).

Lemay Teaches computer science in the LEA department of Université de Lille for around 200h per year (Licence and Master). He is also responsible for computer science and numeric correspondent for its department.

Monet Teaches computer science as a temporary lecturer for a total of 85h per year – 69h for the computer science department of Université de Lille (FIL), and 16h for the computer science department of Centrale Lille. That includes Compilers (M1, 24h), Introduction to Security (18h), Advanced Databases (M1, 27h), Databases 1 (M1, 8h), SQL and Databases (G3 Centrale Lille, 8h).

Niehren Gives lessons for the 2nd year students of the Master Machine Learning (Université de Lille): on Logical foundations of databases (21h).

Paperman Teaches computer science for around 200h per year. He gives lectures for the computer science and math departments. Topics includes an Advance Databases lecture (M2, 24hx3), Algorithmic and Programming in L2 MIASHS (24h), Web Programming in L3 MIASHS and Introduction to Web Technology in L1 SESI (32h).

Salvati Teaches computer science for a total of around 230h per year in computer science departement of Université de Lille. That includes Introduction to Computer Science (L1, 50h), Logic (L3, 50h), Algorithmic and operational research (L3, 36h), Functional Programming (L3, 35h), Research Option (L3, 10h), Semantic Web (M2, 30h), Advanced Databases (M1, 20h).

Staworko Teaches computer science for a total of around 200h in the MIME department (Université de Lille).

Tison Teaches computer science for a total of around 120h at the Université de Lille. That includes Advanced Algorithms and Complexity (42h, Master), Databases (60h, L2), and Logic (18h, L2).

11.2.3 Supervision

Al Serhali PhD project started 2020. On hyperstream programming. Supervised by Niehren.

Barloy PhD project started 2021. On circuits and lower complexity bounds. Supervised by Paperman and Salvati.

Besnier Thomas Intern as part of the "immersion into research" course of Centralle Lille (2nd year student), for about 90h of work for the student during the April-June period. Supervised par Monet.

Crosetti PhD project in progress since 2018. Linear programs with database queries. Supervised by Tison, Capelli, Niehren. With Ramon from Inria Magnet.

Gallot PhD project defense in december 2021. On safety of data transformations. Supervised by Salvati and Lemay.

Soyez-Martin PhD project started 2020. On Streaming with vectors and circuits. Supervised by Salvati and Paperman.

11.2.4 Juries

Lemay Member of the PhD committee of Paul Gallot (Inria Lille), December 16th.

Salvati Member of the PhD committee of Paul Gallot (Inria Lille), December 16th.

Tison Member of the PhD committee of Alain Riveiro (Centrale Lille), January 26th.

Tison Member of the PhD committee of Quentin Mayolle (Centrale Lille), July 13th.

Tison Member of the PhD committee of Paul Gallot (Inria Lille), December 16th.

Tison Member of the selection jury for the "Programme Jeunes Talents" of "For Women in Science Programme" (Loreal Foundation).

Tison Reviewer for a position as Professor for "Knowledge Based Systems" at TU Dortmund University.

Tison Member of the selection committees for: MCF Université d'ARTOIS, MCF Université of Gustave Eiffel, MCF Université de Paris, and Pr Université de la Réunion (2021).

11.3 Popularization

11.3.1 Education

Tison Membre of the guidance and scientific counselling board of Xperium and of the organising committee for the Xperium challenge 2021.

12 Scientific production

12.1 Major publications

[1] A. Amarilli, L. Jachiet and C. Paperman. 'Dynamic Membership for Regular Languages'. In: ICALP. Vol. 48. International Colloquium on Automata, Languages, and Programming (ICALP 2021). Glasgow, Scotland, France, 2nd July 2021, 116:1–116:17. DOI: 10.4230/LIPIcs.ICALP.2021.116. URL: https://hal.archives-ouvertes.fr/hal-03466453.

[2] M. Arenas, P. Barceló, L. Bertossi and M. Monet. 'The Tractability of SHAP-Score-Based Explanations over Deterministic and Decomposable Boolean Circuits'. In: *Thirty-Fifth AAAI Conference on Artificial Intelligence*. Held online, France, Feb. 2021. URL: https://hal.inria.fr/hal-03147623.

[3] C. Barloy, F. Murlak and C. Paperman. 'Stackless Processing of Streamed Trees'. In: *2021 PODS*. Xi'an, Shaanx, China, June 2021. DOI: 10.4230/LIPIcs. URL: https://hal.archives-ouvertes.fr/hal-03021960.

[4] I. Boneva, J. G. Labra Gayo and E. G. Prud 'hommeaux. 'Semantics and Validation of Shapes Schemas for RDF'. In: *ISWC2017 - 16th International semantic web conference*. Vienna, Austria, Oct. 2017. URL: https://hal.archives-ouvertes.fr/hal-01590350.

[5] P. Bourhis, M. Leclère, M.-L. Mugnier, S. Tison, F. Ulliana and L. Gallois. 'Oblivious and Semi-Oblivious Boundedness for Existential Rules'. In: *IJCAI 2019 - International Joint Conference on Artificial Intelligence*. Macao, China, Aug. 2019. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02148142.

[6] F. Capelli, N. Crosetti, J. Niehren and J. Ramon. 'Linear Programs with Conjunctive Queries'. In: 25th Internationcal Conference on Database Theory (ICDT 2022). Edinburgh, United Kingdom, 29th Mar. 2022. URL: https://hal.archives-ouvertes.fr/hal-01981553.

[7] F. Capelli, J.-M. Lagniez and P. Marquis. 'Certifying Top-Down Decision-DNNF Compilers'. In: *Thirty-Fifth AAAI Conference on Artificial Intelligence*. Online, France, Feb. 2021. URL: https://hal.inria.fr/hal-03111679.

[8] P. D. Gallot, A. Lemay and S. Salvati. 'Linear high-order deterministic tree transducers with regular look-ahead'. In: *MFCS 2020 : The 45th International Symposium on Mathematical Foundations of Computer Science*. Andreas Feldmann, Michal Koucky and Anna Kotesovcova. Prague, Czech Republic, Aug. 2020. DOI: 10.4230/LIPIcs.MFCS.2020.34. URL: https://hal.archives-ouvertes.fr/hal-02902853.

[9] J. Niehren and M. Sakho. 'Determinization and Minimization of Automata for Nested Words Revisited'. In: *Algorithms* (Feb. 2021). URL: https://hal.inria.fr/hal-03134596.

[10] S. Staworko and P. Wieczorek. 'Containment of Shape Expression Schemas for RDF'. In: *SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*. Amsterdam, Netherlands, June 2019. URL: https://hal.inria.fr/hal-01959143.

12.2 Publications of the year

International journals

[11] E. Allart, J. Niehren and C. Versari. 'Computing Difference Abstractions of Linear Equation Systems'. In: *Theoretical Computer Science* (20th July 2021). DOI: 10.1016/j.tcs.2021.06.030. URL: https://hal.archives-ouvertes.fr/hal-03156136.

[12] E. Allart, J. Niehren and C. Versari. 'Exact Boolean Abstraction of Linear Equation Systems'. In: *Computation* 9.11 (11th Oct. 2021), p. 32. URL: https://hal.inria.fr/hal-03384058.

[13] M. Arenas, P. Barceló and M. Monet. 'The Complexity of Counting Problems Over Incomplete Databases'. In: *ACM Transactions on Computational Logic* 22.4 (31st Oct. 2021), pp. 1–52. DOI: 10.1145/3461642. URL: https://hal.archives-ouvertes.fr/hal-03356951.

[14] I. Boneva, J. Niehren and M. Sakho. 'Regular Matching and Inclusion on Compressed Tree Patterns with Constrained Context Variables'. In: *Information and Computation* (2021). DOI: 10.1016/j.ic.2021.104776. URL: https://hal.inria.fr/hal-03151014.

[15] A. Bonifati, S. Dumbrava, G. Fletcher, J. Hidders, M. Hofer, W. Martens, F. Murlak, J. Shinavier, S. Staworko and D. Tomaszuk. 'Threshold Queries in Theory and in the Wild'. In: *Proceedings of the VLDB Endowment (PVLDB)* (2022). URL: https://hal.inria.fr/hal-03516360.

[16] M. Cadilhac, F. Mazowiecki, C. Paperman, M. Pilipczuk and G. Sénizergues. 'On Polynomial Recursive Sequences'. In: *Theory of Computing Systems* (2nd June 2021). DOI: 10.1007/s00224-021-10046-9. URL: https://hal.archives-ouvertes.fr/hal-03467171.

[17] J. S. Guez, F. Coucheney, J. Castéra-Guy, M. Béchet, P. Fontanille, N.-E. Chihib, J. Niehren, F. Coutte and P. Jacques. 'Bioinformatics modelling and metabolic engineering of thebranched chain amino acid pathway for specific production of mycosubtilin isoforms in Bacillus subtilis'. In: *Metabolites* 12.2 (19th Jan. 2022). DOI: 10.3390/metabo12020107. URL: https://hal.inria.fr/hal-03498125.

[18] J. Niehren and M. Sakho. 'Determinization and Minimization of Automata for Nested Words Revisited'. In: *Algorithms* (24th Feb. 2021). DOI: 10.3390/a14030068. URL: https://hal.inria.fr/hal-03134596.

[19] A. Théatre, C. Cano-Prieto, M. Bartolini, Y. Laurin, M. Deleu, J. Niehren, T. Fida, S. Gerbinet, M. Alanjary, M. H. Medema, A. Léonard, L. Lins, A. Arabolaza, H. Gramajo, H. Gross and P. Jacques. 'The surfactin-like lipopeptides from Bacillus spp.: natural biodiversity and synthetic biology for a broader application range'. In: *Frontiers in Bioengineering and Biotechnology* (2nd Mar. 2021). DOI: 10.3389/fbioe.2021.623701. URL: https://hal.inria.fr/hal-03158419.

International peer-reviewed conferences

[20] A. Amarilli, L. Jachiet and C. Paperman. 'Dynamic Membership for Regular Languages'. In: ICALP. Vol. 48. International Colloquium on Automata, Languages, and Programming (ICALP 2021). Glasgow, Scotland, France, 2nd July 2021, 116:1–116:17. DOI: 10.4230/LIPIcs.ICALP.2021.116. URL: https://hal.archives-ouvertes.fr/hal-03466453.

[21] R. Angles, A. Bonifati, S. Dumbrava, G. Fletcher, K. W. Hare, J. Hidders, V. E. Lee, B. Li, L. Libkin, W. Martens, F. Murlak, J. Perryman, O. Savković, M. Schmidt, J. Sequeda, S. Staworko and D. Tomaszuk. 'PG-Keys: Keys for Property Graphs'. In: ACM Special Interest Group on Management of Data (SIGMOD). Xi'an, China, June 2021. DOI: 10.1145/3448016.3457561. URL: https://hal.inria.fr/hal-03189192.

[22] M. Arenas, P. Barceló, L. Bertossi and M. Monet. 'The Tractability of SHAP-Score-Based Explanations over Deterministic and Decomposable Boolean Circuits'. In: AAAI 2021 - 35th Conference on Artificial Intelligence. Virtual, France, 2nd Feb. 2021. URL: https://hal.inria.fr/hal-03147623.

[23] C. Barloy, F. Murlak and C. Paperman. 'Stackless Processing of Streamed Trees'. In: PODS 2021 - Symposium on Principles of Database Systems. Proceedings of the Symposium on Principles of Database Systems, PODS 2021. Xi'an, Shaanx, China, 20th June 2021. DOI: 10.4230/LIPIcs. URL: https://hal.archives-ouvertes.fr/hal-03021960.

[24] F. Capelli, N. Crosetti, J. Niehren and J. Ramon. 'Linear Programs with Conjunctive Queries'. In: 25th Internationcal Conference on Database Theory (ICDT 2022). Edinburgh, United Kingdom, 29th Mar. 2022. URL: https://hal.archives-ouvertes.fr/hal-01981553.

[25] F. Capelli, J.-M. Lagniez and P. Marquis. 'Certifying Top-Down Decision-DNNF Compilers'. In: AAAI 2021 - 35th Conference on Artificial Intelligence. Virtual, France, 2nd Feb. 2021. URL: https://hal.inria.fr/hal-03111679.

[26] D. Deutch, N. Frost, B. Kimelfeld and M. Monet. 'Computing the Shapley Value of Facts in Query Answering'. In: SIGMOD Conference 2022. Philadelphia, United States, 12th June 2022. URL: https://hal.inria.fr/hal-03514297.

[27] B. Groz, A. Lemay, S. Staworko and P. Wieczorek. 'Inference of Shape Graphs for Graph Databases'. In: International Conference on Database Theory. Edinbourgh, United Kingdom, 2022. DOI: 10.4230/LIPIcs.ICDT.2022.7. URL: https://hal.inria.fr/hal-03559309.

Doctoral dissertations and habilitation theses

[28] P. D. G. Gallot. 'Safety of transformations of data trees: Tree transducer theory applied to a verification problem on shell scripts'. Université de Lille, 16th Dec. 2021. URL: https://hal.archives-ouvertes.fr/tel-03517128.

Reports & preprints

[29] E. Allart, J. Niehren and C. Versari. *Reaction Networks to Boolean Networks*. 15th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-02279942.

[30] I. Boneva, L. Staworko and J. Lozano. *Consistency and Certain Answers in Relational to RDF Data Exchange with Shape Constraints*. 10th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03474916.

[31] M. Cadilhac and C. Paperman. *The Regular Languages of Wire Linear AC 0*. 8th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03466451.

[32] J. Niehren, M. Sakho and A. Al Serhali. *Schema-Based Automata Determinization*. 19th Jan. 2022. URL: https://hal.inria.fr/hal-03536045.