

Report from Dagstuhl Seminar 24032

Representation, Provenance, and Explanations in Database Theory and Logic

Pablo Barcelo^{*1}, Pierre Bourhis^{*2}, Stefan Mengel^{*3}, and Sudeepa Roy^{*4}

- 1 PUC – Santiago de Chile, CL. pbarcelo@ing.puc.cl
- 2 CNRS – CRIStAL, Lille, FR. pierre.bourhis@univ-lille.fr
- 3 CNRS, CRIL – Lens, FR. mengel@cril.fr
- 4 Duke University – Durham, US. sudeepa@cs.duke.edu

Abstract

This report documents the program and the outcomes of **Dagstuhl Seminar “Representation, Provenance, and Explanations in Database Theory and Logic” (24032)**, which was broadly in the area of database theory. Database theory formalizes the theoretical underpinnings of databases and analyzes them with mathematical tools. We focused on questions related to the fundamental problem of efficient query evaluation: compute the answers of a query on a database. This seminar focused on three key aspects of query evaluations. (1) **Representation** studies the tradeoff between expressivity, compactness, and efficient computation of outputs from the inputs, including circuits and knowledge compilation forms, enumeration, and direct access. (2) **Provenance** captures the computation process of outputs from the inputs using a compact formula, and has applications to probabilistic databases. (3) **Explanations** give meaningful insights to responsibilities of different inputs toward an output beyond provenance, e.g., by using Shapley Values from co-operative game theory that has been recently popular in both DB and ML.

Seminar January 14–19, 2024 – <https://www.dagstuhl.de/24032>

2012 ACM Subject Classification Theory of computation → Theory and algorithms for application domains; Theory of computation → Theory and algorithms for application domains; Theory of computation → Theory and algorithms for application domains; Theory of computation → Theory and algorithms for application domains; Theory of computation → Theory and algorithms for application domains

Keywords and phrases Circuits, database theory, factorized databases, provenance, shapley values

Digital Object Identifier 10.4230/DagRep.14.1.49

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Representation, Provenance, and Explanations in Database Theory and Logic, *Dagstuhl Reports*, Vol. 14, Issue 1, pp. 49–71

Editors: Pablo Barcelo, Pierre Bourhis, Stefan Mengel, and Sudeepa Roy



DAGSTUHL Dagstuhl Reports

REPORTS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany


1 Executive Summary

Pablo Barcelo (PUC – Santiago de Chile, CL)

Pierre Bourhis (CNRS – CRISTAL, Lille, FR)

Stefan Mengel (CNRS, CRIL – Lens, FR)

Sudeepa Roy (Duke University – Durham, US)

License  Creative Commons BY 4.0 International license
 © Pablo Barcelo, Pierre Bourhis, Stefan Mengel, and Sudeepa Roy

Background and Research area

The Dagstuhl Seminar “*Representation, Provenance, and Explanations in Database Theory and Logic*” (24032) was broadly in *Database Theory*, where the goal is to formalize the theoretical underpinnings of databases and then analyze them with mathematical tools. One of the most fundamental problems in both database theory and systems is efficient query evaluation: given a database and a query, compute the answer to the query on the database. This question has a tight connection to logic, since it has been known for a long time that different fragments of first- or second-order logic can be seen as the core of practical query languages like SQL or Datalog. This seminar focused on three key aspects of query evaluations: *representation*, *provenance*, and *explanations*.

Representation. For large datasets, query results can be very large when they are materialized explicitly in the standard form. For efficient query processing and subsequent applications, it is important to *represent* the query answers in a compact fashion. One important form of representations in query evaluation is by *circuits*, which have a long history in complexity theory and AI and can be seen as part of the larger framework of *knowledge compilation* (Darwiche, Marquis, J. Artif. Intell. Res. 2002). Circuits were heavily discussed in several presentations in the seminar. The other aspect of representation that the seminar focused on was the field of *enumeration algorithms* and *direct access*. It first computes a data structure representing the query answers, and then gives an algorithm to extract one answer at a time from the data structure. In this problem, the complexity of the two parts is measured separately: the computation time of the data structure is called the *preprocessing time* and the time of the extraction of each answer is called the *delay*. Typically, the goal of such algorithms is to have a preprocessing time much smaller than the cost of the classical evaluation of the query and very small (ideally constant) delay.

Provenance. Data provenance in general refers to how the outputs of a query are generated from the inputs, with a broad goal to enable interpretability, trust, and reproducibility of the queries. A mathematical form of provenance that propagates annotations of inputs to the outputs, called *provenance semirings*, was proposed in a seminal work by Green et al. (PODS 2007). The most specialized case of Boolean semirings captures how an output tuple has been obtained from the inputs with joint usage (joins – translate to conjunctions \wedge), and alternative usages (projections or unions – translate to disjunctions \vee). Such semirings can be used to understand compactly how outputs are generated from inputs, and have applications in *query evaluation in probabilistic databases* when realization of inputs tuples is uncertain (Dalvi-Suciu, JACM 2012), and in *deletion propagation* or *view update*, to understand how the outputs change if one or more inputs are deleted, without re-computing the query. There are more advanced semirings like tropical semirings that can capture shortest paths in graphs. Compact and efficient knowledge compilations of provenance circuits into *ordered and free binary decision diagrams (OBDDs, FBDDs)*, and more generally as *decomposable*

deterministic negation normal forms (d-DNNF) are also important questions in database theory with applications in probabilistic databases (Jha-Suciu, ICDT 2011; Beame et al., ACM Trans. Database Syst. 2017; Monet, PODS 2020).

Explanations. While provenance provides one approach to explaining query answers capturing how the query answers are generated, in many applications, other forms of insights as explanations are desired for understanding contributions of inputs, trends and anomalies in the outputs, and deciding next course of actions or recourse. Recently, explanations based on the widely known *Shapley values* from co-operative game theory have been used in database theory to measure the relevance of a certain database fact to a query answer (Deutch et al, SIGMOD 2022; Livshits et al., ICDT 2021), and to measure the relevance of inputs to the outcome of an ML classifier (Arenas et al., AAAI 2021). Complexity, applications, and algorithms for explanations by Shapley values were heavily discussed in the seminar. Since the naive computation of Shapley values is intractable as it includes a summation over exponentially many subsets, one of the main themes behind this investigation has been the identification of practically relevant classes of database queries for which such explanations can be computed in polynomial time, possibly using knowledge compilation forms. Apart from Shapley values, other forms of explanations, including that of aggregated database query answers (e.g., Roy-Suciu, SIGMOD'14) and connections of explanations with data privacy, fairness, and causal inference were discussed in the seminar. This way the seminar connected the field of database theory to the field of *responsible data science* that is of paramount importance in real world.

Acknowledgements

We are grateful to the Scientific Directorate and to the staff of the Schloss Dagstuhl – Leibniz Center for Informatics for their support of this seminar.

2 Table of Contents

Executive Summary	
<i>Pablo Barcelo, Pierre Bourhis, Stefan Mengel, and Sudeepa Roy</i>	50
Organization of the Seminar	54
Outcomes of the seminar	54
Overview of Talks	56
Consistency of Relations over Monoids	
<i>Albert Atserias</i>	56
Lower Bounds on Probabilistic Query Evaluation	
<i>Antoine Amarilli</i>	56
Privately Generating Justifiably Fair Data	
<i>Amir Gilad</i>	57
Model Interpretability through the Lens of Computational Complexity	
<i>Pablo Barcelo</i>	57
SHAP-Scores and Its Computation over ML Models	
<i>Pablo Barcelo</i>	58
A dichotomy for succinct representations of homomorphisms	
<i>Christoph Berkholz</i>	58
Tractability and Optimization of Shap-Score Computation for Explainable AI	
<i>Leopoldo Bertossi</i>	59
Circuits for Query Evaluation over Trees	
<i>Pierre Bourhis</i>	59
From Queries to Circuits	
<i>Florent Capelli</i>	60
Unified Reverse Data Management	
<i>Wolfgang Gatterbauer</i>	60
Graph Explainability and Shapley	
<i>Floris Geerts</i>	61
Lessons Learned from Building Systems for Provenance and Explanations	
<i>Boris Glavic</i>	61
Answering Database Queries Using Direct-Access Structures	
<i>Benny Kimelfeld</i>	62
The Shapley Value in Database Management	
<i>Ester Livshits</i>	62
Provenance in Queries, Games, and Argumentation: Time for a Family Reunion	
<i>Bertram Ludäscher</i>	63
Impact of Self-Joins on Enumeration and Direct Access on Join Queries	
<i>Stefan Mengel</i>	64
The Intensional-Extensional Problem in Probabilistic Databases	
<i>Mikaël Monet</i>	64

Revisiting Semiring Provenance for Datalog <i>Liat Peterfreund</i>	65
Datalog over (Pre-)Semirings <i>Reinhard Pichler</i>	65
MSO Enumeration over Words and their Representations <i>Cristian Riveros</i>	66
Explanations for Aggregate Query Answers – An Overview <i>Sudeepa Roy</i>	67
Training Invariant Machine Learning Models with Incomplete Data <i>Babak Salimi</i>	67
Expected Shapley-Like Scores of Boolean Functions: Complexity and Applications to Probabilistic Databases <i>Pierre Senellart</i>	68
The Importance of Parameters in Database Queries <i>Christoph Standke</i>	68
From Shapley Value to Model Counting and Back <i>Dan Suciu</i>	69
Answering Quantile Join Queries by Representing Inequality Predicates Efficiently <i>Nikolaos Tziavelis</i>	69
Open problems	70
A problem on unambiguous DNFs <i>Mikaël Monet</i>	70
A question about representability of probabilistic databases <i>Christoph Standke</i>	70
Participants	71

3 Organization of the Seminar

The seminar was held between January 14–19, 2024 (Monday to Friday with arrival on Sunday). We had 26 on-site participants. We started the first day with an introduction of each participant presenting their background, research area, as well as what they wished to achieve from the seminar. Right after, we had the opening keynote (the only one-hour talk) of the seminar by Reinhard Pichler on Datalog over semirings. We had a mix of 45 mins, 30 mins, and 20 mins talks in the rest of the seminar. On the first day, we had 8 talks of different length, focusing on backgrounds on semirings, explanations for database queries and explanations in ML, and Shapley values, and short talks on various topics later in the day. The aim was to cover a significant part of the background for the rest of the seminar as well as to learn about interesting research from several of the participants on the very first day. This allowed us to have more relaxed schedule in the rest of the week with more time for free collaboration, as well as to schedule more technical talks later in the week. On Tuesday and Wednesday, we had talks on model counting, probabilistic databases, enumeration, direct access, semirings, and circuits. On Thursday, we focused on systems and application aspects, including causal inference, fairness, and privacy, and short talks on miscellaneous topics. We had ample time of free discussions from Tuesday to Friday (including the typical time for excursion on Wednesday afternoon, which had to be canceled because of bad weather). We saw talks ranging from logic and complexity, systems, to applications related to the seminar topics. There were 26 talks spread over the first four days of the seminar given by 25 participants, and two open problem sessions (Thursday morning and Friday morning). All talks were well received, with many questions and lively discussions during and after the talks. Overall, the seminar was highly engaging, intellectually stimulating, and a great success.

4 Outcomes of the seminar

- Scientific content:** The participants learnt about backgrounds and recent work on the seminar topics – *representations, provenance, and explanations*, from experts. In the opening keynote, Reinhard Pichler gave a comprehensive introduction to *provenance semirings* that was used in a large number of talks in the seminar. He also talked about their recent work on the query language Datalogo, which is based on the concept of K-relations and generalizes recursive Datalog to (pre-)semirings. Later, we saw talks on different semantics of provenance semirings for Datalog (Liat Peterfreund), consistency of relations over monoids (Albert Atserias), and provenance in queries, games, and argumentation (Bertram Ludascher).

We saw different views and applications of *explanations* in the seminar. Sudeepa Roy talked about explanations for aggregate query answers: in response to user questions on why an output is high/low, or higher/lower than other attributes, how to generate deep explanations that the domain experts can find from the data automatically. Pablo Barcelo talked about another form of explanations, a framework for judging and comparing the interpretability of different ML models, and complexity of this problem. A number of presentations discussed various aspects of *Shapley (SHAP) values* as explanations: computing SHAP scores over ML models (Pablo Barcelo), using Shapley values to measure the responsibility of individual database tuples to the outcome for query answering and database inconsistency (Ester Livshits), tractability of SHAP scores for explainable AI (Leo Bertossi), use of Shapley-like scores for explaining graph neural networks (Floris

Geerts), polynomial-time equivalence between computation of Shapley values and model counting for a class of functions (Dan Suciu), equivalent tractability of the computations of expected Shapley values and of the expected values of Boolean functions in probabilistic databases (Pierre Senellart), and quantifying the importance of the choices of parameter values to the result of a query over a database using SHAP scores (Christoph Standke). We had several talks on *query evaluations on probabilistic (uncertain) data*, which made connections between provenance polynomials and their representations as circuits. Antoine Amarilli talked about lower bounds for probabilistic query evaluation, for both computation and size of provenance as circuits. Mikael Monet revisited the intensional-extensional problem in probabilistic databases, and talked about their ongoing work on whether the tractability for UCQ can be captured by knowledge compilation. Pierre Bourhis talked about circuits for query evaluation over trees, and Florent Capelli discussed algorithms to construct tractable circuits from queries.

For *representations* we had multiple talks on direct access and enumeration. Stefan Mengel talked about the impact of self-join for such queries. Cristian Riveros presented a survey of MSO enumeration problems over words based on the model of annotated automata. Benny Kimelfeld discussed fine-grained complexity of database queries that involve joins, grouping, aggregation, and ordering using direct access structures. Nikos Tziavelis talked about the complexity of answering quantile join queries by efficiently representing inequality predicates.

Wolfgang Gatterbauer made a connection between the problem of finding minimal size *provenance factorizations* and reverse data management problems such as resilience (how to change a query answer with smallest change in data). Christoph Berkholz discussed lower bounds for factorized representations for multi-way join queries and homomorphisms between two structures.

On the *systems and applications* side, Boris Glavic shared with the participants the lessons he learned from his work on building systems for capturing and managing provenance and explanations, and the separation between data flow between operators in a query and the information flow by provenance. Making connections with responsible data science, Babak Salimi discussed the challenges and solutions in training ML models with incomplete data and in the presence of selection bias in data, and its applications in the context of fairness. Amir Gilad presented a framework for synthetic data generation that is both differentially-private and fair.

- **Open problems:** The participants discussed several open problems in the open problem sessions. For instance, (1) what are the notions equivalent to semirings for datalog with negation? Earlier, monus has been proposed for non-monotone queries. How do the requirements for being stable (from the recent work on Datalogo over semirings) and having a monus interact? (2) How do we define and complexity for Shapley values for queries with negation and queries with aggregates? Do the approximation results from the recent literature still hold when we have negation? For queries with aggregates, should we assign responsibilities to single tuples or a group of tuples? Two other open problems are listed at the end of this document.
- **Making connections between seminar topics and theory, systems, and applications:** The seminar brought together researchers who are broadly interested in one or more of the seminar topics, but work on different aspects of these topics. While a majority of the participants work on the theoretical aspects of the topics, some participants work on systems and the other work on applications and data science. We also saw interesting exchanges of ideas among different topics (representations, provenance, and explanations) in the seminar.

- **Extensive collaborations:** In addition to learning about recent research from the talks, the participants extensively discussed problems with old or new collaborators during the week.

5 Overview of Talks

5.1 Consistency of Relations over Monoids

Albert Atserias (UPC Barcelona Tech, ES)

License © Creative Commons BY 4.0 International license
© Albert Atserias

Joint work of Albert Atserias, Phokion G. Kolaitis

Main reference Albert Atserias, Phokion G. Kolaitis: “Consistency of Relations over Monoids”, CoRR, Vol. abs/2312.02023, 2023.

URL <https://doi.org/10.48550/ARXIV.2312.02023>

The interplay between local consistency and global consistency has been the object of study in several different areas, including probability theory, relational databases, and quantum information. For relational databases, Beeri, Fagin, Maier, and Yannakakis showed that a database schema is acyclic if and only if it has the local-to-global consistency property for relations, which means that every collection of pairwise consistent relations over the schema is globally consistent. More recently, the same result has been shown under bag semantics. In this paper, we carry out a systematic study of local vs. global consistency for relations over positive commutative monoids, which is a common generalization of ordinary relations and bags. Let K be an arbitrary positive commutative monoid. We begin by showing that acyclicity of the schema is a necessary condition for the local-to-global consistency property for K -relations to hold. Unlike the case of ordinary relations and bags, however, we show that acyclicity is not always sufficient. After this, we characterize the positive commutative monoids for which acyclicity is both necessary and sufficient for the local-to-global consistency property to hold; this characterization involves a combinatorial property of monoids, which we call the *transportation property*. We then identify several different classes of monoids that possess the transportation property. As our final contribution, we introduce a modified notion of local consistency of K -relations, which we call *pairwise consistency up to the free cover*. We prove that, for all positive commutative monoids K , even those without the transportation property, acyclicity is both necessary and sufficient for every family of K -relations that is pairwise consistent up to the free cover to be globally consistent.

5.2 Lower Bounds on Probabilistic Query Evaluation

Antoine Amarilli (Telecom Paris, FR)

License © Creative Commons BY 4.0 International license
© Antoine Amarilli

This talk focuses on the task of computing the probability that a fixed query holds on an input probabilistic database. The problem can also be specialized to several contexts, e.g., computing the probability that an input graph with probabilistic edges contains a specific pattern, or in the unweighted case counting how many subgraphs of the input have a certain property. We will review recent hardness results on this problem. We will cover two kinds of

results: lower bounds on the computational complexity of the problem, and lower bounds on the size of the query provenance when represented in structured circuit classes.

References

- 1 Antoine Amarilli, Timothy van Bremen, Kuldeep S. Meel: Conjunctive Queries on Probabilistic Graphs: The Limits of Approximability. ICDT 2024.
- 2 Antoine Amarilli. Uniform Reliability for Unbounded Homomorphism-Closed Graph Queries. ICDT 2023.
- 3 Antoine Amarilli, Benny Kimelfeld. Uniform Reliability of Self-Join-Free Conjunctive Queries. LMCS, 2022.
- 4 Antoine Amarilli, Mikaël Monet. Weighted Counting of Matchings in Unbounded-Treewidth Graph Families. MFCS 2022.

5.3 Privately Generating Justifiably Fair Data

Amir Gilad (The Hebrew University of Jerusalem, IL)

License © Creative Commons BY 4.0 International license
© Amir Gilad

Joint work of David Pujol, Amir Gilad, Ashwin Machanavajjhala

Main reference David Pujol, Amir Gilad, Ashwin Machanavajjhala: “PreFair: Privately Generating Justifiably Fair Synthetic Data”, Proc. VLDB Endow., Vol. 16(6), pp. 1573–1586, 2023.

URL <https://doi.org/10.14778/3583140.3583168>

In this talk, I will present our recent work that develops a framework for synthetic data generation that is both differentially-private and fair, where fairness is modeled by an adaptation of the causal definition for justifiable fairness.

5.4 Model Interpretability through the Lens of Computational Complexity

Pablo Barcelo (PUC – Santiago de Chile, CL)

License © Creative Commons BY 4.0 International license
© Pablo Barcelo

Joint work of Pablo Barcelo, Bernardo Subercaseaux


This talk revisits a framework for judging and comparing the interpretability of classes of Machine Learning models. Said framework allows us to formalize and prove a nuanced version of claims like “decision trees are more interpretable than neural networks”. Interestingly, such a formalization pointed out the first result establishing the hardness of interpreting decision trees, and provided tools to analyze how hyper-parameters such as the number of layers in a network can impact its interpretability. Our framework relied on a few assumptions that will be discussed explicitly in the talk, such as the role of well-defined interpretability queries or the adequacy of computational complexity for capturing the practical complexity of real-life instances.

References

- 1 Model Interpretability through the Lens of Computational Complexity. Pablo Barceló, Mikaël Monet, Jorge Pérez, Bernardo Subercaseaux. (<https://arxiv.org/abs/2010.12265>)
- 2 On Computing Probabilistic Explanations for Decision Trees. Marcelo Arenas, Pablo Barceló, Miguel Romero, Bernardo Subercaseaux. (<https://arxiv.org/abs/2207.12213>)

5.5 SHAP-Scores and Its Computation over ML Models

Pablo Barcelo (PUC – Santiago de Chile, CL)

License  Creative Commons BY 4.0 International license
© Pablo Barcelo

Main reference Marcelo Arenas, Pablo Barceló, Leopoldo E. Bertossi, Mikaël Monet: “On the Complexity of SHAP-Score-Based Explanations: Tractability via Knowledge Compilation and Non-Approximability Results”, *J. Mach. Learn. Res.*, Vol. 24, pp. 63:1–63:58, 2023.

URL <http://jmlr.org/papers/v24/21-0389.html>

SHAP scores are expressions designed to capture the contribution of a feature to the output of a machine learning model. They are grounded in the well-studied game-theoretical notion of Shapley values. In this discussion, I will elucidate the meaning of these SHAP scores expressions and explain how they are obtained from first principles. Subsequently, I will delve into the examination of the problem of computing SHAP scores over machine learning models. I will provide insights into when and why this problem becomes computationally intractable. Additionally, I will identify a large and practically relevant class of models for which the problem can be solved in polynomial time. Finally, I will show that even for slight extensions of this class, the computation of SHAP scores is not only intractable but also does not admit a Fully Polynomial Randomized Approximation Scheme (FPRAS).

5.6 A dichotomy for succinct representations of homomorphisms

Christoph Berkholz (TU Ilmenau, DE)

License  Creative Commons BY 4.0 International license
© Christoph Berkholz

Main reference Christoph Berkholz, Harry Vinnal-Smeeth: “A Dichotomy for Succinct Representations of Homomorphisms”, in *Proc. of the 50th International Colloquium on Automata, Languages, and Programming, ICALP 2023, July 10-14, 2023, Paderborn, Germany, LIPIcs, Vol. 261*, pp. 113:1–113:19, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

URL <https://doi.org/10.4230/LIPICS.ICALP.2023.113>

The talk is based on the cited ICALP’23 paper. It will be about factorized databases for multi-way join queries, or, in other words, succinct representations of all homomorphisms between two structures A and B. The main result is a characterisation of (bounded-arity) structures A where this is efficiently doable. In the talk I will mainly focus on lower bounds for factorized representations.

5.7 Tractability and Optimization of Shap-Score Computation for Explainable AI

Leopoldo Bertossi (SKEMA Business School – Montréal, CA)

License © Creative Commons BY 4.0 International license
© Leopoldo Bertossi

Joint work of Marcelo Arenas, Pablo Barceló, Mikäel Monet, Jorge León

Main reference Leopoldo E. Bertossi, Jorge E. León: “Efficient Computation of Shap Explanation Scores for Neural Network Classifiers via Knowledge Compilation”, in Proc. of the Logics in Artificial Intelligence – 18th European Conference, JELIA 2023, Dresden, Germany, September 20-22, 2023, Proceedings, Lecture Notes in Computer Science, Vol. 14281, pp. 49–64, Springer, 2023.

URL https://doi.org/10.1007/978-3-031-43619-2_4

The presentation is about recent research on the Shap Scores in Explainable Machine Learning. More specifically, on the basis of the tractability result for Shap [1] for open-box classifiers defined by a class of Boolean circuits (actually, d-DBC), we show how Shap can be computed much more efficiently than through the sheer use of the classifier’s input/output relation when a Binary Neural Network classifier is, first, represented by means of a compact CNF formula, which is, next, (knowledge) compiled into an SDD, followed by a transformation into a d-DBC [2].

References

- 1 Marcelo Arenas, Pablo Barcelo, Leopoldo Bertossi, Mikäel Monet. “On the Complexity of SHAP-Score-Based Explanations: Tractability via Knowledge Compilation and Non-Approximability Results”. *Journal of Machine Learning Research*, 2023, 24(63):1-58.
- 2 Leopoldo Bertossi and Jorge E. León. “Efficient Computation of Shap Explanation Scores for Neural Network Classifiers via Knowledge Compilation”. Proc. of JELIA’23, Springer LNCS 14281, 2023, pp. 49-64.

5.8 Circuits for Query Evaluation over Trees

Pierre Bourhis (CNRS – CRISAL, Lille, FR)

License © Creative Commons BY 4.0 International license
© Pierre Bourhis

Querying trees via Tree automata presents a lot of interest because several important questions can be executed with a guaranteed efficient time. Over the last decades, different approaches have been presented to solve major query answering questions such as enumeration, probabilistic evaluation... In this survey, we review a particular approach which can be adapted to all these questions: a knowledge compilation approach. We present the different results that can be resolved by this approach and also its limits.

References

- 1 Antoine Amarilli, Pierre Bourhis, Florent Capelli, Mikäel Monet: Ranked Enumeration for MSO on Trees via Knowledge Compilation. CoRR abs/2310.00731 (2023)
- 2 Antoine Amarilli, Pierre Bourhis, Stefan Mengel, Matthias Niewerth: Enumeration on Trees with Tractable Combined Complexity and Efficient Updates. PODS 2019: 89-103
- 3 Antoine Amarilli, Pierre Bourhis, Stefan Mengel: Enumeration on Trees under Relabelings. ICDT 2018: 5:1-5:18
- 4 Antoine Amarilli, Pierre Bourhis, Louis Jachiet, Stefan Mengel: A Circuit-Based Approach to Efficient Enumeration. ICALP 2017: 111:1-111:15
- 5 Antoine Amarilli, Pierre Bourhis, Pierre Senellart: Provenance Circuits for Trees and Treelike Instances. ICALP (2) 2015: 56-68

5.9 From Queries to Circuits

Florent Capelli (University of Artois/CNRS – Lens, FR)

License © Creative Commons BY 4.0 International license
© Florent Capelli

Main reference Florent Capelli, Oliver Irwin: “Direct Access for Conjunctive Queries with Negations”, in Proc. of the 27th International Conference on Database Theory, ICDT 2024, March 25-28, 2024, Paestum, Italy, LIPIcs, Vol. 290, pp. 13:1–13:20, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024.

URL <https://doi.org/10.4230/LIPICS.ICDT.2024.13>

In this talk, we will review two algorithms to construct tractable circuits from a conjunctive query and a database whose size can be bounded using fractional hypertree width. The first one is a classical bottom up dynamic programming on a join tree of the conjunctive query, which can be seen as a generalization of Yannakakis approach. The second one is based on a top-down approach akin to exhaustive DPLL, an algorithm originally devised for solving #SAT. We will show that both algorithms construct very similar circuits on conjunctive queries but that DPLL can be applied to a more general setting without changing much of its structure.

5.10 Unified Reverse Data Management

Wolfgang Gatterbauer (Northeastern University – Boston, US)

License © Creative Commons BY 4.0 International license
© Wolfgang Gatterbauer

Joint work of Wolfgang Gatterbauer, Neha Makhija

What is a minimal set of tuples to delete from a database in order to eliminate all query answers? This problem is called “the resilience of a query” and is one of the key algorithmic problems underlying various forms of reverse data management, such as view maintenance, deletion propagation and causal responsibility. A long-open question is determining the conjunctive queries (CQs) for which resilience can be solved in PTIME.

We shed new light on this problem by proposing a unified Integer Linear Programming (ILP) formulation. It is unified in that it can solve both previously studied restrictions (e.g., self-join-free CQs under set semantics that allow a PTIME solution) and new cases (all CQs under set or bag semantics). It is also unified in that all queries and all database instances are treated with the same approach, yet the algorithm is guaranteed to terminate in PTIME for all known PTIME cases. In particular, we prove that for all known easy cases, the optimal solution to our ILP is identical to a simpler Linear Programming (LP) relaxation, which implies that standard ILP solvers return the optimal solution to the original ILP in PTIME.

In broader terms, we believe that using one single algorithm that can solve all queries (easy and hard) and then proving that it terminates in PTIME for the subset of PTIME queries will become a conventional and unified approach for attacking several other open problems in reverse data management.

References

- 1 Makhija and Gatterbauer. “A Unified Approach for Resilience and Causal Responsibility with Integer Linear Programming (ILP) and LP Relaxations”, SIGMOD 2024. <https://dl.acm.org/doi/pdf/10.1145/3626715>
- 2 Makhija and Gatterbauer. “Towards a Dichotomy for Minimally Factorizing the Provenance of Self-Join Free Conjunctive Queries”, PODS 2024. <https://arxiv.org/pdf/2105.14307>

5.11 Graph Explainability and Shapley

Floris Geerts (University of Antwerp, BE)

License © Creative Commons BY 4.0 International license
© Floris Geerts

Main reference Shichang Zhang, Yozen Liu, Neil Shah, Yizhou Sun: “GStarX: Explaining Graph Neural Networks with Structure-Aware Cooperative Games”, in Proc. of the Advances in Neural Information Processing Systems, Vol. 35, pp. 19810–19823, Curran Associates, Inc., 2022.

URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7d53575463291ea6b5a23cf6e571f59b-Paper-Conference.pdf

Graph explainability is a critical aspect in understanding and interpreting complex relationships within graph-structured data. The need for transparent and interpretable models has led to the exploration of various methodologies, with a focus on providing insights into the contribution of individual nodes or edges in a graph. Shapley values, inspired by cooperative game theory, offer a principled approach to attribute values to each node, reflecting their marginal contributions to different coalitions. Myerson value further refines this concept by considering the externalities of a coalition, providing a more comprehensive understanding of node importance. In the context of graph explainability, Hamiache and Navarro score introduces a novel perspective by evaluating the relevance of nodes based on the information flow and connectivity patterns, offering a nuanced interpretation of their impact on the overall graph structure. Together, these approaches contribute to the development of explainable graph models, enabling stakeholders to gain deeper insights into the dynamics and significance of individual elements within complex graph data.

5.12 Lessons Learned from Building Systems for Provenance and Explanations

Boris Glavic (University of Illinois – Chicago, US)

License © Creative Commons BY 4.0 International license
© Boris Glavic

In this talk I will introduce to the audience lessons learned from my work and other group’s work on building systems for capturing and managing provenance and explanations. For instance, an underappreciated concept in developing such systems is that provenance creates a separate information flow in the system that does not conform to the standard way of how data flows through the operators of a query.

References

- 1 <https://vldb.org/pvldb/vol115/p451-niu.pdf>
- 2 <https://arxiv.org/pdf/1804.07156.pdf>
- 3 <http://sites.computer.org/debull/A18mar/p51.pdf>
- 4 <http://www.vldb.org/pvldb/vol113/p912-lee.pdf>
- 5 <https://dl.acm.org/doi/pdf/10.1145/3555041.3589731>
- 6 <https://inria.hal.science/hal-01851538/document>
- 7 <https://dl.acm.org/doi/pdf/10.14778/2824032.2824089>

5.13 Answering Database Queries Using Direct-Access Structures

Benny Kimelfeld (Technion – Haifa, IL)

License  Creative Commons BY 4.0 International license
© Benny Kimelfeld


The talk will describe recent results on the fine-grained complexity of database queries that involve joins, grouping, aggregation, and ordering. For some common aggregate functions (e.g., min, max, count, sum), such a query can be phrased as an ordinary conjunctive query over a database annotated with a suitable commutative semiring. I will discuss the ability to evaluate such queries by constructing, in quasilinear time in the database size (i.e., roughly the time it takes to read the database), a data structure that provides logarithmic-time direct access to the answers, ordered by a desired lexicographic order. This task is nontrivial since the number of answers might be larger than quasilinear in the database size, so, the data structure needs to provide a representation that is compact, easy to construct, and fast to access. The results provide classifications of queries, orderings, and semirings by the feasibility of such complexity guarantees.

References

- 1 Nofar Carmeli, Nikolaos Tziavelis, Wolfgang Gatterbauer, Benny Kimelfeld, Mirek Riedewald: Tractable Orders for Direct Access to Ranked Answers of Conjunctive Queries. *PODS 2021*: 325-341
- 2 Idan Eldar, Nofar Carmeli, Benny Kimelfeld: Direct Access for Answers to Conjunctive Queries with Aggregation. *CoRR abs/2303.05327 (2023)*. *ICDT 2024*.

5.14 The Shapley Value in Database Management

Ester Livshits (University of Edinburgh, GB)

License  Creative Commons BY 4.0 International license
© Ester Livshits

Joint work of Ester Livshits, Leopoldo E. Bertossi, Benny Kimelfeld, Moshe Sebag
Main reference Ester Livshits, Leopoldo E. Bertossi, Benny Kimelfeld, Moshe Sebag: “The Shapley Value of Tuples in Query Answering”, *Log. Methods Comput. Sci.*, Vol. 17(3), 2021.
URL [https://doi.org/10.46298/LMCS-17\(3:22\)2021](https://doi.org/10.46298/LMCS-17(3:22)2021)

We consider two situations where we wish to quantify the responsibility of individual database tuples to the outcome. The first is query answering, where we wish to provide an explanation as to why we obtained a specific answer. The second is database inconsistency, where the goal is to identify the most problematic tuples. Some tuples may contribute more than others to the outcome, which can be a bit in the case of a Boolean query, a tuple or a number for conjunctive and aggregate queries, respectively, or a number indicating how inconsistent the database is (i.e., an inconsistency measure). To quantify the contribution of tuples, we use the well-known Shapley value that was introduced in cooperative game theory in the 1950s and has found applications in a plethora of domains. We investigate the applicability of the Shapley value in the two settings, as well as the computational aspects of its calculation in terms of complexity, algorithms, and approximation.

References

- 1 Ester Livshits, Leopoldo E. Bertossi, Benny Kimelfeld, and Moshe Sebag. “The Shapley Value of Tuples in Query Answering”. *Logical Methods in Computer Science* (2021). <https://lmcs.episciences.org/8437>

- 2 Ester Livshits and Benny Kimelfeld. “The Shapley Value of Inconsistency Measures for Functional Dependencies”. *Logical Methods in Computer Science* (2022). <https://lmcs.episciences.org/9705>

5.15 Provenance in Queries, Games, and Argumentation: Time for a Family Reunion

Bertram Ludäscher (University of Illinois at Urbana-Champaign, US)

License © Creative Commons BY 4.0 International license
© Bertram Ludäscher

Joint work of Bertram Ludäscher, Shawn Bowers, Yilin Xia

Main reference Bertram Ludäscher, Shawn Bowers, Yilin Xia: “Games, Queries, and Argumentation Frameworks: Towards a Family Reunion”, in Proc. of the 7th Workshop on Advances in Argumentation in Artificial Intelligence (AI³ 2023) co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AIXIA 2023), Rome, Italy, November 9, 2023, CEUR Workshop Proceedings, Vol. 3546, CEUR-WS.org, 2023.

URL <https://ceur-ws.org/Vol-3546/paper06.pdf>

Consider the non-stratified, recursive query Q : $\text{win}(X) \text{ :- move}(X,Y)$, not $\text{win}(Y)$.

Its 2-valued reading states that a position x in a two-player game is won if there exists a move to a position y that is lost (not won). If the move graph contains cycles, drawn positions may occur (neither player can force a win). It is well known that the 3-valued well-founded semantics can be used to solve games: $\text{win}(x)$ is true, false, and undefined, respectively, iff x is won, lost, or drawn.

The query Q has been used in logic programming (e.g., to illustrate the well-founded semantics [3], in database theory (e.g., to show that stratified Datalog is strictly less expressive than the class of Fixpoint queries [2], and in formal argumentation (as a meta-interpreter for abstract argumentation frameworks).

Solved game graphs can be said to “explain themselves” (or contain their own provenance “for free”): The provenance of a won, lost, or drawn position is easily obtained via an RPQ-definable subgraph of the solved (labeled) game graph in which positions and moves have an associated value (won, lost, or drawn for positions, and winning, delaying, drawing, or blundering for moves, respectively). Since Q is a syntactic variant of Dung’s meta-interpreter for abstract argument frameworks AF [4], the provenance structure available in solved game graphs can be used to explain and justify the grounded (i.e., well-founded) extensions of AF. Another application of game provenance are query evaluation games: The n -ary version of Q can be understood as a normal form for Fixpoint, i.e., all Fixpoint queries can be rewritten into a game normal form, even when restricted to draw-free games [5]. For the subclass of FO queries (First-Order queries expressed in Datalog syntax), this normal form has been used to derive an elegant and powerful provenance representation that unifies how-provenance and why-not provenance [6].

References

- 1 B. Ludäscher, S. Bowers, and Y. Xia. Games, Queries, and Argumentation Frameworks: Towards a Family Reunion. *AI³@AI*IA* (2023)
- 2 P. Kolaitis, The expressive power of stratified logic programs, *Information and Computation* (1991).
- 3 A. Van Gelder, K. A. Ross, J. S. Schlipf, The Well-founded Semantics for General Logic Programs, *Journal of the ACM* (1991).
- 4 P. Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, *Logic Programming and n-Person Games, AI* (1995)

- 5 J. Flum, M. Kubierschky, and B. Ludäscher. Total and partial well-founded Datalog coincide. ICDT, Delphi. LNCS 1186 (1997)
- 6 S. Köhler, B. Ludäscher, and D. Zinn. First-Order Provenance Games. In Search of Elegance in the Theory and Practice of Computation (Peter Buneman Festschrift). LNCS 8000 (2013)

5.16 Impact of Self-Joins on Enumeration and Direct Access on Join Queries

Stefan Mengel (CNRS, CRIL – Lens, FR)

License  Creative Commons BY 4.0 International license
 © Stefan Mengel

Joint work of Karl Bringmann, Nofar Carmeli, Stefan Mengel, Luc Segoufin

It has been known essentially since the introduction of conjunctive queries that self-joins have an impact on the evaluation of join queries. While in settings like answering Boolean queries and counting their complexity implications are completely understood, the situation is far less clear for other query answering tasks. In this talk, I will present some recent progress for enumeration (Carmeli and Segoufin 2023) and direct access (Bringmann, Carmeli, and Mengel 2023) showing that, even though these settings are often conceptually very close, self-joins behave very differently for them.

References

- 1 Karl Bringmann, Nofar Carmeli, Stefan Mengel: Tight Fine-Grained Bounds for Direct Access on Join Queries. CoRR abs/2201.02401 (2022)
- 2 Nofar Carmeli, Luc Segoufin: Conjunctive Queries With Self-Joins, Towards a Fine-Grained Enumeration Complexity Analysis. PODS 2023: 277-289

5.17 The Intensional-Extensional Problem in Probabilistic Databases

Mikaël Monet (INRIA Lille, FR)

License  Creative Commons BY 4.0 International license
 © Mikaël Monet

Joint work of Mikaël Monet, Antoine Amarilli, Dan Suciu

Main reference Mikaël Monet: “Solving a Special Case of the Intensional vs Extensional Conjecture in Probabilistic Databases”, CoRR, Vol. abs/1912.11864, 2019.

URL <http://arxiv.org/abs/1912.11864>

Dalvi and Suciu established a dichotomy for probabilistic query evaluation (PQE) over tuple-independent databases, for unions of conjunctive queries (UCQs): for each UCQ, the problem is either solvable in PTIME, or is #P-hard. The UCQs for which the problem is in PTIME are called **safe**. Dalvi and Suciu’s algorithm on such a safe query relies essentially on the following three probabilistic rules: Independence, Negation, and Inclusion-Exclusion. In parallel, another method to obtain PTIME algorithms for PQE is through **knowledge compilation**: one first compiles the provenance of a query Q on a TID D as a Boolean circuit or diagram from the field of knowledge compilation (e.g., OBDDs, FBDDs, d-DNNFs, etc), and then uses this circuit to compute the probability. At a high-level, this type of algorithm makes use of the following three probabilistic rules: Independence, Negation, and **Disjoint union** (instead of inclusion-exclusion). This naturally leads to the following question, called the intensional-extensional problem: letting Q be a safe UCQ, can the tractability of PQE(Q) be captured with the knowledge compilation approach?

In this talk I will talk about this problem, present a technique that allowed to handle a specific class of UCQs, and discuss our ongoing work on the problem. In particular, I will present a neat combinatorial conjecture, that we named the “non-cancelling intersections” conjecture, that talks only about sets and the so-called Möbius function (i.e., no databases, no queries, no complexity). This talk is based on ongoing work with Antoine Amarilli, Louis Jachiet, and Dan Suciu.

References

- 1 Abhay Kumar Jha, Dan Suciu: Knowledge Compilation Meets Database Theory: Compiling Queries to Decision Diagrams. *Theory Comput. Syst.* 52(3): 403-440 (2013)
- 2 Michaël Monet: Solving a Special Case of the Intensional vs Extensional Conjecture in Probabilistic Databases. *PODS 2020*: 149-163

5.18 Revisiting Semiring Provenance for Datalog

Liat Peterfreund (The Hebrew University of Jerusalem, IL)

License © Creative Commons BY 4.0 International license
© Liat Peterfreund

Joint work of Camille Bourgaux, Pierre Bourhis, Liat Peterfreund, Michaël Thomazo

Main reference Camille Bourgaux, Pierre Bourhis, Liat Peterfreund, Michaël Thomazo: “Revisiting Semiring Provenance for Datalog”, in *Proc. of the 19th International Conference on Principles of Knowledge Representation and Reasoning*, pp. 91–101, 2022.

URL <https://doi.org/10.24963/kr.2022/10>

While the definition of semiring provenance is uncontroversial for unions of conjunctive queries, the picture is less clear for Datalog. Indeed, the original definition might include infinite computations and is not consistent with other proposals for Datalog semantics over annotated data. In this work, we propose and investigate several provenance semantics, based on different approaches for defining classical Datalog semantics. We study the relationship between these semantics, and introduce properties that allow us to analyze and compare them.

5.19 Datalog over (Pre-)Semirings

Reinhard Pichler (TU Wien, AT)

License © Creative Commons BY 4.0 International license
© Reinhard Pichler

Joint work of Mahmoud Abo Khamis, Hung Q. Ngo, Reinhard Pichler, Dan Suciu, Yisu Remy Wang

Main reference Mahmoud Abo Khamis, Hung Q. Ngo, Reinhard Pichler, Dan Suciu, Yisu Remy Wang: “Convergence of Datalog over (Pre-) Semirings”, in *Proc. of the PODS ’22: International Conference on Management of Data*, Philadelphia, PA, USA, June 12 – 17, 2022, pp. 105–117, ACM, 2022.

URL <https://doi.org/10.1145/3517804.3524140>

Datalog is a successful query language that extends relational calculus by recursion, has an elegant declarative semantics as well as a simple operational semantics, and admits several powerful optimizations such as semi-naïve evaluation and magic set rewriting. However, datalog also has its limitations since it only supports monotone queries over sets. This means, for instance, that aggregates (which are crucial in many data analytics tasks but are not monotone under set inclusion) are not supported in pure datalog.

In a seminal paper by Green, Karvounarakis, and Tannen [1], K-relations were introduced as a generalization of standard relations. In a K-relation, tuples are mapped to some semiring K. We can then consider standard relations as K-relations over the Boolean semiring, bags

of tuples as K-relations over the natural numbers, sparse tensors as K-relations over the reals, etc. Also provenance information at various levels of detail can be captured by an appropriate choice of the semiring K.

In this talk, I have presented our recent work [2, 3] on the query language datalogo, which is based on the concept of K-relations and generalizes datalog to (pre-)semirings. In particular, I have shown how it can capture various computations involving aggregates as well as provenance information. Moreover, I have briefly mentioned convergence properties of datalogo and some optimization techniques.

References

- 1 Todd J. Green, Gregory Karvounarakis, Val Tannen: Provenance semirings. PODS 2007: 31-40: <https://dl.acm.org/doi/10.1145/1265530.1265535>.
- 2 Mahmoud Abo Khamis, Hung Q. Ngo, Reinhard Pichler, Dan Suciu, Yisu Remy Wang: Convergence of Datalog over (Pre-) Semirings. PODS 2022: 105-117: <https://dl.acm.org/doi/10.1145/3517804.3524140>, full version (to appear in J.ACM): <https://arxiv.org/abs/2105.14435>.
- 3 Yisu Remy Wang, Mahmoud Abo Khamis, Hung Q. Ngo, Reinhard Pichler, Dan Suciu: Optimizing Recursive Queries with Program Synthesis. SIGMOD Conference 2022: 79-93: <https://dl.acm.org/doi/10.1145/3514221.3517827>.

5.20 MSO Enumeration over Words and their Representations

Cristian Riveros (PUC – Santiago de Chile, CL)

License  Creative Commons BY 4.0 International license
© Cristian Riveros

I will present a survey of MSO enumeration problems over words based on the model of annotated automata, a model for encoding MSO queries with output. In the first half, I will present the basic MSO enumeration problem and the representations needed for efficient enumeration. In the second half, I will go through extensions of this MSO enumeration problem, with the required extensions on the representations. Toward the end, I will present some open problems.

References

- 1 Martín Muñoz, Cristian Riveros: Constant-Delay Enumeration for SLP-Compressed Documents. ICDT 2023.
- 2 Martín Muñoz, Cristian Riveros: Streaming Enumeration on Nested Documents. ICDT 2022.
- 3 Antoine Amarilli, Pierre Bourhis, Louis Jachiet, Stefan Mengel: A Circuit-Based Approach to Efficient Enumeration. ICALP 2017.

5.21 Explanations for Aggregate Query Answers – An Overview

Sudeepa Roy (Duke University – Durham, US)

License © Creative Commons BY 4.0 International license
© Sudeepa Roy

Joint work of Michael Cafarella, Sainyam Galhotra, Boris Glavic, Amir Gilad, Chenjie Li, Zhengjie Miao, Sudeepa Roy, Babak Salimi, Dan Suciu, Brit Youngmann, Qitian Zeng

I will give an overview of different types of explanations for aggregate query answers answering user questions like why a value is high/low or higher/lower than another value. I will discuss explanations by intervention, counterbalance, augmented provenance, causal explanations, and actionable explanations. Explanations by Shapley Value will be covered in other talks.

References

- 1 Sudeepa Roy, Dan Suciu: A Formal Approach to Finding Explanations for Database Queries, SIGMOD 2014
- 2 Zhengjie Miao, Qitian Zeng, Boris Glavic, Sudeepa Roy: Going Beyond Provenance: Explaining Query Answers with Pattern-based Counterbalances, SIGMOD 2019
- 3 Chenjie Li, Zhengjie Miao, Qitian Zeng, Boris Glavic, Sudeepa Roy: Putting Things into Context: Rich Explanations for Query Answers using Join Graphs, SIGMOD 2021
- 4 Sainyam Galhotra, Amir Gilad, Sudeepa Roy, Babak Salimi: Hyper: Hypothetical Reasoning With What-If and How-To Queries Using a Probabilistic Causal Approach, SIGMOD 2022
- 5 Brit Youngmann, Amir Gilad, and Michael Cafarella, Sudeepa Roy: Summarized Causal Explanations For Aggregate Views, SIGMOD 2024

5.22 Training Invariant Machine Learning Models with Incomplete Data

Babak Salimi (University of California, San Diego – La Jolla, US)

License © Creative Commons BY 4.0 International license
© Babak Salimi

Main reference Jiongli Zhu, Sainyam Galhotra, Nazanin Sabri, Babak Salimi: “Consistent Range Approximation for Fair Predictive Modeling”, Proc. VLDB Endow., Vol. 16(11), pp. 2925–2938, 2023.

URL <https://doi.org/10.14778/3611479.3611498>

In this talk, I aim to discuss the significant challenge of learning machine learning models that satisfy invariant properties under conditional independence constraints. The importance of this problem will be illustrated through various real-world examples, emphasizing its relevance and urgency. Subsequently, I will analyze existing approaches and their shortcomings, especially in situations where data is compromised by quality issues such as selection bias. To overcome these obstacles, I will introduce a framework inspired by techniques for querying incomplete data in data management. This framework is tailored to effectively handle the specific challenges posed by incomplete datasets. Additionally, I will demonstrate its application in the context of algorithmic fairness.

5.23 Expected Shapley-Like Scores of Boolean Functions: Complexity and Applications to Probabilistic Databases

Pierre Senellart (ENS, PSL University – Paris, FR)

License © Creative Commons BY 4.0 International license
© Pierre Senellart

Joint work of Pratik Karmakar, Mikaël Monet, Pierre Senellart, Stephane Bressan

Main reference Pratik Karmakar, Mikaël Monet, Pierre Senellart, Stephane Bressan: “Expected Shapley-Like Scores of Boolean functions: Complexity and Applications to Probabilistic Databases”, Proc. ACM Manag. Data, Vol. 2(2), Association for Computing Machinery, 2024.

URL <https://doi.org/10.1145/3651593>

Shapley values, originating in game theory and increasingly prominent in explainable AI, have been proposed to assess the contribution of facts in query answering over databases, along with other similar power indices such as Banzhaf values. In this work we adapt these Shapley-like scores to probabilistic settings, the objective being to compute their expected value. We show that the computations of expected Shapley values and of the expected values of Boolean functions are interreducible in polynomial time, thus obtaining the same tractability landscape. We investigate the specific tractable case where Boolean functions are represented as deterministic decomposable circuits, designing a polynomial-time algorithm for this setting. We present applications to probabilistic databases through database provenance, and an effective implementation of this algorithm within the ProVSQL system, which experimentally validates its feasibility over a standard benchmark.

References

- 1 Pratik Karmakar, Mikaël Monet, Pierre Senellart, and Stéphane Bressan, 2024. Expected Shapley-Like Scores of Boolean Functions: Complexity and Applications to Probabilistic Databases, <https://arxiv.org/abs/2401.06493>
- 2 Daniel Deutch, Nave Frost, Benny Kimelfeld, and Mikaël Monet. 2022. Computing the Shapley value of facts in query answering. In SIGMOD Conference. ACM, 1570–1583.
- 3 Pierre Senellart, Louis Jachiet, Silviu Maniu, and Yann Ramusat. 2018. ProVSQL: Provenance and Probability Management in PostgreSQL. Proc. VLDB Endow. 11, 12 (2018), 2034–2037.

5.24 The Importance of Parameters in Database Queries

Christoph Standke (RWTH Aachen, DE)

License © Creative Commons BY 4.0 International license
© Christoph Standke

Joint work of Martin Grohe, Benny Kimelfeld, Peter Lindner, Christoph Standke

Main reference Martin Grohe, Benny Kimelfeld, Peter Lindner, Christoph Standke: “The Importance of Parameters in Database Queries”, in Proc. of the 27th International Conference on Database Theory, ICDT 2024, March 25-28, 2024, Paestum, Italy, LIPIcs, Vol. 290, pp. 14:1–14:17, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024.

URL <https://doi.org/10.4230/LIPICS.ICDT.2024.14>

In this talk, I will introduce a framework for quantifying the importance of the choices of parameter values to the result of a query over a database. In our framework, the importance of a parameter is its SHAP score and we make the case for the rationale of using this score by showing that we arrive at this score in two different, apparently opposing, approaches to quantifying the contribution of a parameter. We then point out that this framework yields an interesting complexity-theoretic landscape.

5.25 From Shapley Value to Model Counting and Back

Dan Suciu (*University of Washington – Seattle, US*)

License  Creative Commons BY 4.0 International license
© Dan Suciu

We study the problem of quantifying the contribution of each Boolean variable to the satisfying assignments of a Boolean function, based on the Shapley value. This problem was introduced by Livshits et al. in order to quantify the contribution of an input tuple to the output of a query. We prove polynomial-time equivalence between computing Shapley values and model counting, for classes of Boolean functions that are closed under substitutions of variables with disjunctions of fresh variables. This result settles an open problem raised by Deutch et al., which sought to connect the Shapley value computation to probabilistic query evaluation.

References

- 1 Ester Livshits, Leopoldo E. Bertossi, Benny Kimelfeld, Moshe Sebag: The Shapley Value of Tuples in Query Answering. *Log. Methods Comput. Sci.* 17(3) (2021)
- 2 Daniel Deutch, Nave Frost, Benny Kimelfeld, Mikaël Monet: Computing the Shapley Value of Facts in Query Answering. *SIGMOD Conference 2022*: 1570-1583
- 3 Ahmet Kara, Dan Olteanu, Dan Suciu: From Shapley Value to Model Counting and Back. *CoRR abs/2306.14211* (2023) (To appear in *PODS'2024*)

5.26 Answering Quantile Join Queries by Representing Inequality Predicates Efficiently

Nikolaos Tziavelis (*Northeastern University – Boston, US*)

License  Creative Commons BY 4.0 International license
© Nikolaos Tziavelis


Joint work of Nikolaos Tziavelis, Nofar Carmeli, Wolfgang Gatterbauer, Benny Kimelfeld, Mirek Riedewald
Main reference Nikolaos Tziavelis, Nofar Carmeli, Wolfgang Gatterbauer, Benny Kimelfeld, Mirek Riedewald: “Efficient Computation of Quantiles over Joins”, in *Proc. of the 42nd ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2023, Seattle, WA, USA, June 18-23, 2023*, pp. 303–315, ACM, 2023.
URL <https://doi.org/10.1145/3584372.3588670>


We consider the complexity of answering Quantile Join Queries, which ask for the answer at a specified relative position (e.g., 50% for the median) under some ordering over the answers to an ordinary Join Query (JQ). Compared to the task of direct access, this task is easier since only one access is required. The goal is to avoid materializing the set of all join answers, and to achieve quasilinear time in the size of the database, regardless of the total number of answers. The tractability of such a query does not only depend on the join structure, but also on the desired order. We show an algorithm that covers all known tractable cases by iteratively using a “trimming” subroutine which removes query answers that are higher or lower (according to the ranking function) than a certain answer determined as the “pivot”. Trimming essentially adds inequality predicates to our initial query and an efficient representation of these inequalities implies efficient Quantile Join Query answering for a large class of ranking functions.

6 Open problems

6.1 A problem on unambiguous DNFs

Mikaël Monet (INRIA Lille, FR)

License  Creative Commons BY 4.0 International license

 Mikaël Monet


URL <https://csttheory.stackexchange.com/q/53733>

I presented the problem that can be found here: <https://csttheory.stackexchange.com/q/53733>.

6.2 A question about representability of probabilistic databases

Christoph Standke (RWTH Aachen, DE)

License  Creative Commons BY 4.0 International license

 Christoph Standke

Joint work of Christoph Standke, Peter Lindner, Dan Suciu, Dan Olteanu, Christopher Ré, Christoph Koch
Main reference Dan Suciu, Dan Olteanu, Christopher Ré, Christoph Koch: “Probabilistic Databases”, Morgan & Claypool Publishers, 2011.

URL <https://doi.org/10.2200/S00362ED1V01Y201105DTM016>

Given a finite probabilistic database as a set of instance-probability pairs, (how) can we decide whether this probabilistic database can be obtained via a finite tuple-independent probabilistic database and a view consisting of conjunctive queries?

Participants

- Antoine Amarilli
Telecom Paris, FR
- Albert Atserias
UPC Barcelona Tech, ES
- Pablo Barcelo
PUC – Santiago de Chile, CL
- Christoph Berkholz
TU Ilmenau, DE
- Leopoldo Bertossi
SKEMA Business School –
Montréal, CA
- Pierre Bourhis
CNRS – CRISStAL, Lille, FR
- Florent Capelli
University of Artois/CNRS –
Lens, FR
- Wolfgang Gatterbauer
Northeastern University –
Boston, US
- Floris Geerts
University of Antwerp, BE
- Amir Gilad
The Hebrew University of
Jerusalem, IL
- Boris Glavic
University of Illinois –
Chicago, US
- Benny Kimelfeld
Technion – Haifa, IL
- Ester Livshits
University of Edinburgh, GB
- Bertram Ludäscher
University of Illinois at
Urbana-Champaign, US
- Stefan Mengel
CNRS, CRIL – Lens, FR
- Mikaël Monet
INRIA Lille, FR
- Liat Peterfreund
The Hebrew University of
Jerusalem, IL
- Reinhard Pichler
TU Wien, AT
- Cristian Riveros
PUC – Santiago de Chile, CL
- Sudeepa Roy
Duke University – Durham, US
- Babak Salimi
University of California,
San Diego – La Jolla, US
- Pierre Senellart
ENS, PSL University – Paris, FR
- Christoph Standke
RWTH Aachen, DE
- Dan Suciu
University of Washington –
Seattle, US
- Nikolaos Tziavelis
Northeastern University –
Boston, US
- Harry Vinall-Smeeth
TU Ilmenau, DE

