DAGSTUHL
REPORTS

Volume 9, Issue 9, September 2019

Dagstuhl Reports, Vol. 9, Issue 9 ISSN 2192-5283

ISSN 2192-5283

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at
http://www.dagstuhl.de/dagpub/2192-5283

Publication date
March, 2020

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at `http://dnb.d-nb.de`.

Digital Object Identifier: 10.4230/DagRep.9.9.i

Aims and Scope
The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.
In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e. g. summaries from panel discussions or open problem sessions.

Editorial Office
Michael Wagner*(Managing Editor)*
Jutka Gasiorowski *(Editorial Assistance)*
Dagmar Glaser *(Editorial Assistance)*
Thomas Schillo *(Technical Assistance)*

Contact
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de
http://www.dagstuhl.de/dagrep

Report from Dagstuhl Seminar 19361

Logic and Learning

Edited by

Michael Benedikt[1], Kristian Kersting[2], Phokion G. Kolaitis[3], and Daniel Neider[4]

1 **University of Oxford, GB,** michael.benedikt@gmail.com
2 **TU Darmstadt, DE,** kersting@cs.tu-darmstadt.de
3 **University of California – Santa Cruz and**
IBM Almaden Research Center – San Jose, US, kolaitis@soe.ucsc.edu
4 **MPI-SWS – Kaiserslautern, DE,** neider@mpi-sws.org

Abstract

The goal of building truly intelligent systems has forever been a central problem in computer science. While logic-based approaches of yore have had their successes and failures, the era of machine learning, specifically deep learning is also coming upon significant challenges. There is a growing consensus that the inductive reasoning and complex, high-dimensional pattern recognition capabilities of deep learning models need to be combined with symbolic (even programmatic), deductive capabilities traditionally developed in the logic and automated reasoning communities in order to achieve the next step towards building intelligent systems, including making progress at the frontier of hard problems such as explainable AI. However, these communities tend to be quite separate and interact only minimally, often at odds with each other upon the subject of the “correct approach” to AI. This report documents the efforts of Dagstuhl Seminar 19361 on “Logic and Learning” to bring these communities together in order to: (i) bridge the research efforts between them and foster an exchange of ideas in order to create unified formalisms and approaches that bear the advantages of both research methodologies; (ii) review and analyse the progress made across both communities; (iii) understand the subtleties and difficulties involved in solving hard problems using both perspectives; (iv) make attempts towards a consensus on what the hard problems are and what the elements of good solutions to these problems would be.

The three focal points of the seminar were the strands of “Logic for Machine Learning”, “Machine Learning for Logic”, and “Logic vs. Machine Learning”. The seminar format consisted of long and short talks, as well as breakout sessions. We summarise the motivations and proceedings of the seminar, and report on the abstracts of the talks and the results of the breakout sessions.

Seminar September 1–6, 2019 – http://www.dagstuhl.de/19361

2012 ACM Subject Classification Computing methodologies → Artificial intelligence, Computing methodologies → Machine learning approaches, Computing methodologies → Logical and relational learning, Theory of computation → Constraint and logic programming, Theory of computation → Models of learning, General and reference → Surveys and overviews, Theory of computation → Logic

Keywords and phrases Artificial Intelligence, Automated reasoning, Databases, Deep Learning, Inductive Logic Programming, Logic, Logic and Learning, Logic for Machine Learning, Logic vs. Machine Learning, Machine Learning, Machine Learning for Logic, Neurosymbolic methods

Digital Object Identifier 10.4230/DagRep.9.9.1

Edited in cooperation with Adithya Murali

Logic and Learning, *Dagstuhl Reports*, Vol. 9, Issue 9, pp. 1–22
Editors: Michael Benedikt, Kristian Kersting, Phokion G. Kolaitis, and Daniel Neider
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Michael Benedikt (University of Oxford, GB)
Kristian Kersting (TU Darmstadt, DE)
Phokion G. Kolaitis (University of California – Santa Cruz & IBM Almaden Research Center – San Jose, US)
Adithya Murali (University of Illinois – Urbana-Champaign, US)
Daniel Neider (MPI-SWS – Kaiserslautern, DE)

Motivation

Logic and learning are central to Computer Science, and in particular to AI research and allied areas. Alan Turing envisioned, in his paper "Computing Machinery and Intelligence" [1], a combination of statistical (*ab initio*) machine learning and an "unemotional" symbolic language such as logic. However, currently, the interaction between research in logic and research in learning is far too limited; in fact, they are often perceived as being completely distinct or even opposing approaches.

While there has been interest in using machine learning methods within many application areas of logic, the investigation of these interactions has usually been carried out within the confines of a single problem area. We believe that an interaction involving a broader perspective is needed. It would be fruitful to look for common techniques in applying learning to logic-related tasks, which requires looking across a wide spectrum of applications. It is also important to consider the ways that logic and learning, deduction and induction, can work together.

Design of the Seminar

The main aim of this Dagstuhl Seminar was to address the above problems by bring researchers from the logic and learning communities together and to create bridges between the two fields via the exchange of ideas ranging between the (seemingly) polar possibilities of the injection of declarative methods in machine learning and the use and applications of learning technologies in logical contexts. This included creating an understanding of the work in different applications, an increased understanding of the formal connections between these applications, and the development of a more unified view of the current attempts to organically reconcile deductive and inductive approaches. In order to structure these explorations, the focal points of the seminar were the following three distinct strands of interaction between logic and learning:

1. *Machine Learning for Logic*, including the learning of logical artifacts, such as formulas, logic programs, database queries and integrity constraints, as well as the application of learning to tune deductive systems.
2. *Logic for Machine Learning*, including the role of logics in delineating the boundary between tractable and intractable learning problems, the construction of formalisms that allow learning systems to take advantage of specified logical rules, and the use of logic as a declarative framework for expressing machine learning constructs.
3. *Logic vs. Machine Learning*, including the study of problems that can be solved using either logic-based techniques or via machine learning, an exploration of the trade-offs between these techniques, and the development of benchmarks for comparing these methods.

Summary of seminar activities

The seminar was attended by 41 researchers across various communities including logic, databases, Inductive Logic Programming (ILP), formal verification, machine learning, deep learning, and theorem proving. The membership consisted of senior and junior researchers, including graduate students, post-doctoral researchers, and industry experts. The seminar was conducted through talks and breakout sessions, with breaks for discussion between the attendees. There were three long talks, 21 short talks, and three breakout sessions on the discussion of open problems in logic and learning.

The talks consisted of: (i) presentation of recent advances in research questions and methodologies relating to the motivations discussed above; (ii) surveys of the state of research on various problems requiring the combination of deductive and inductive reasoning as well as methodologies developed to address fundamental hurdles in this space; (iii) new perspectives on the organic combination of logical formulations and methods with machine learning in specific application domains; (iv) theoretical formulations and results on problems in learning logical representations; (v) demonstrations of state-of-the art tools combining logic and learning for applications such as theorem proving or entity resolution; (vi) presentation of research on challenge problems for the field of AI and intelligent reasoning.

The breakout sessions were conducted in three continuing parts, each spanning one session. The first part involved all the participants in a discussion of the current (small and large) open problems in AI, challenge problems for the field of intelligent systems, and research questions about defining specific goals representing a successful combination of inductive and deductive reasoning. This involved a deliberation of what problems were relevant, which problems could be potentially related to or dependent upon each other, and various suggestions to formalise commonly desired research goals. This session resulted in the choice of three broad areas for further specific discussion: (i) Explainable AI (ii) Injecting symbolic knowledge or constraints into neural networks, and (iii) Learning of logical formulae (first-order logic) from satisfaction on structures in a differentiable manner. The second part consisted of parallel thematic sessions on these three areas. Each thematic session was conducted in the form of a round-table discussion and was led by one or two participants who championed the theme. The third session brought all the participants together again to conclude with a summary of the ideas exchanged during the parallel sessions.

Conclusion

We consider the seminar a success. There is a growing need to enable the disparate communities of logic and learning to interact with each other, and we noted from the seminar that researchers from each community appreciated the perspective offered by the other, often identified techniques used by the other community that could be imported into their own, and, interestingly, were in agreement about the relevant and important problems of the day. The format of the seminar including ample time for discussions and breakout sessions received positive feedback from the participants.

References

1 A. M. Turing, “Computing machinery and intelligence”, Mind, vol. LIX, pp. 433–460, October 1950

2 Table of Contents

Executive Summary

Overview of Talks

Breakout Sessions

3 Overview of Talks

3.1 Six perspectives on logic & learning (in infinite domains)

Vaishak Belle (University of Edinburgh, GB)

The unification of low-level perception and high-level reasoning is a long-standing problem in artificial intelligence, and among other approaches, the integration of logic and learning potentially offers the most general solution to that problem. Although there has been considerable progress on this integration, models in practise continue to make the finite domain assumption, and so models are essentially propositional, programs are loop-free, and so on. In this talk, we discuss a number of different ways in which the infinite is embraced. In recent work, for example, we have looked at the problems of inference and (parameter and structure) learning in continuous domains, that is, where logical atoms model continuous properties. In other work, we report on the synthesis of plans with loops in the presence of probabilistic nondeterminism. Finally, we touch on proposals for declaratively modelling logical reasoning, probabilistic inference and learning problems in continuous domains.

This talk reports on joint work with a number of collaborators and is drawn from the following papers: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11].

References

1. Amelie Levray, Vaishak Belle, "Learning Tractable Probabilistic Models in Open Worlds", CoRR, abs/1901.05847, 2019.
2. Laszlo Treszkai, Vaishak Belle, "A Correctness Result for Synthesizing Plans With Loops in Stochastic Domains", CoRR, abs/1905.07028, 2019.
3. Vaishak Belle, "Abstracting Probabilistic Models", CoRR, abs/1810.02434, 2018.
4. Vaishak Belle, Luc De Raedt, "Semiring Programming: A Framework for Search, Inference and Learning", CoRR, abs/1609.06954, 2016.
5. Vaishak Belle, Brendan Juba, "Implicitly Learning to Reason in First-Order Logic", NeurIPS 2019.
6. Stefanie Speichert, Vaishak Belle, "Learning Probabilistic Logic Programs in Continuous Domains", 29th International Conference on Inductive Logic Programming (ILP), 2019.
7. Andreas Bueff, Stefanie Speichert, Vaishak Belle, "Tractable Querying and Learning in Hybrid Domains via Sum-Product Networks", Workshop on Hybrid Reasoning and Learning (HRL), KR, 2018.
8. Samuel Kolb, Martin Mladenov, Scott Sanner, Vaishak Belle, Kristian Kersting, "Efficient Symbolic Integration for Probabilistic Inference", Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Main track, Pages 5031-5037, July 2018.
9. Vaishak Belle, "Open-Universe Weighted Model Counting", Thirty-First AAAI Conference on Artificial Intelligence, 2017.
10. Davide Nitti, Vaishak Belle, Tinne Laet, Luc De Raedt, Machine Learning, Volume 106 Issue 12, Pages 1905-1932, December 2017.
11. Vaishak Belle, Andrea Passerini, Guy Van Den Broeck, "Probabilistic inference in hybrid domains by weighted model integration", Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15), Pages 2770-2776, July 2015.

3.2 Neural Model Counting

Ismail Ilkan Ceylan (University of Oxford, GB)

Joint work of Ralph Abboud, Ismail Ilkan Ceylan, Thomas Lukasiewicz
Main reference Ralph Abboud, İsmail İlkan Ceylan, Thomas Lukasiewicz: "Learning to Reason: Leveraging Neural Networks for Approximate DNF Counting", CoRR, Vol. abs/1904.02688, 2019.
URL https://arxiv.org/abs/1904.02688

Weighted model counting (WMC) has emerged as a prevalent approach for probabilistic inference. In its most general form, WMC is #P-hard and, as a result, solving real-world WMC instances is intractable. Weighted DNF counting (weighted #DNF) is a special case where approximations with probabilistic guarantees can be tractably obtained, but this requires time O(mn), where m denotes the number of variables and n the number of clauses of the input DNF. In this talk, I will present a novel approach for weighted #DNF that combines approximate model counting with deep learning and accurately approximates model counts in just O(m + n). Our experiments show that our model learns and generalizes very well to large-scale #DNF instances.

3.3 Learning Constraints from Examples

Luc De Raedt (KU Leuven, BE)

Joint work of Luc De Raedt, Andreas Passerini, Stefano Teso
Main reference Luc De Raedt, Andrea Passerini, Stefano Teso: "Learning Constraints From Examples", in Proc. of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pp. 7965–7970, 2018.
URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17229

While constraints are ubiquitous in artificial intelligence and constraints are also commonly used in machine learning and data mining, the problem of learning constraints from examples has received less attention. In this talk I shall discuss the problem of constraint learning in detail, indicate some subtle differences with standard machine learning problems, sketch some applications and summarize the state-of-the-art.

3.4 Query Learning of Omega Regular Languages

Dana Fisman (Ben Gurion University – Beer Sheva, IL)

Joint work of Dana Fisman, Dana Angluin, Timos Antonopoulos, Udi Boker
Main reference Dana Fisman: "Inferring regular languages and ω-languages", J. Log. Algebr. Meth. Program., Vol. 98, pp. 27–49, 2018.
URL https://doi.org/10.1016/j.jlamp.2018.03.002

Omega languages, i.e. languages of infinite words (or of infinite trees), play an important role in modeling, verification and synthesis of reactive systems. While query learning of regular languages of finite words can be done in polynomial time using a polynomial number of membership and equivalence queries, there is no known polynomial learning algorithm for the full class of omega regular languages. In this talk we will discuss the obstacles in obtaining a polynomial learning algorithm and go through state-of-the art results on learning of regular languages of infinite words and of infinite trees.

3.5 Bounds in Query Learning

James Freitag (University of Illinois – Chicago, US)

Joint work of Hunter Chase, James Freitag
Main reference Hunter Chase, James Freitag: "Bounds in Query Learning", CoRR, Vol. abs/1904.10122, 2019.
URL https://arxiv.org/abs/1904.10122

I will discuss some bounds in query learning related to combinatorial quantities isolated in model theory, namely, Littlestone dimension and consistency dimension. These quantities are related to exact learning by equivalence queries and learning by equivalence queries and membership queries. Both quantities were also isolated in model theory (with different names), but can be formulated in a purely combinatorial manner. I will also discuss other potential connections between combinatorial notions from model theory and various settings of learning.

3.6 Learning Logically Specified Problems

Martin Grohe (RWTH Aachen, DE)

Joint work of Martin Grohe, Gyorgy Turan, Martin Ritzert
Main reference Martin Grohe, Martin Ritzert: "Learning first-order definable concepts over structures of small degree", in Proc. of the 32nd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2017, Reykjavik, Iceland, June 20-23, 2017, pp. 1–12, IEEE Computer Society, 2017.
URL https://doi.org/10.1109/LICS.2017.8005080

After some general remarks about learning frameworks for logical specifications, applications scenarios, and practical challenges, I will focus on a declarative model theoretic learning framework. Within this framework, I will talk about recent positive and negative learnability results that we obtained for learning models specified in first-order and monadic second-order logic.

3.7 On Learning to Prove

Daniel Huang (University of California – Berkeley, US)

Main reference Daniel Huang: "On Learning to Prove", CoRR, Vol. abs/1904.11099, 2019.
URL https://arxiv.org/abs/1904.11099

In this talk, we consider the problem of learning a first-order theorem prover that uses a representation of beliefs in mathematical claims to construct proofs. The inspiration for doing so comes from the practices of human mathematicians where"plausible reasoning" is applied in addition to deductive reasoning to find proofs. Towards this end, we introduce a representation of beliefs that assigns probabilities to the exhaustive and mutually exclusive first-order possibilities found in Hintikka's theory of distributive normal forms. The representation supports Bayesian update, induces a distribution on statements that does not enforce that logically equivalent statements are assigned the same probability, and suggests an embedding of statements into an associated Hilbert space. We then examine conjecturing as model selection and an alternating-turn game of determining consistency. The game is amenable (in principle) to self-play training to learn beliefs and derive a prover that is complete when logical omniscience is attained and sound when beliefs are reasonable. The representation has super-exponential space requirements as a function of quantifier depth so the ideas in this paper should be taken as theoretical. We will comment on how abstractions can be used to control the space requirements at the cost of completeness.

3.8 Counterexample-Guided Strategy Improvement for POMDPs Using Recurrent Neural Networks

Nils Jansen (Radboud University Nijmegen, NL)

Joint work of Steven Carr, Nils Jansen, Ralf Wimmer, Alexandru Serban, Bernd Becker, Ufuk Topcu
Main reference Steven Carr, Nils Jansen, Ralf Wimmer, Alexandru Constantin Serban, Bernd Becker, Ufuk Topcu: "Counterexample-Guided Strategy Improvement for POMDPs Using Recurrent Neural Networks", in Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, pp. 5532–5539, ijcai.org, 2019.
URL https://doi.org/10.24963/ijcai.2019/768

We study strategy synthesis for partially observable Markov decision processes (POMDPs). The particular problem is to determine strategies that provably adhere to (probabilistic) temporal logic constraints. This problem is computationally intractable and theoretically hard. We propose a novel method that combines techniques from machine learning and formal verification. First, we train a recurrent neural network (RNN) to encode POMDP strategies. The RNN accounts for memory-based decisions without the need to expand the full belief space of a POMDP. Secondly, we restrict the RNN-based strategy to represent a finite-memory strategy and implement it on a specific POMDP. For the resulting finite Markov chain, efficient formal verification techniques provide provable guarantees against temporal logic specifications. If the specification is not satisfied, counterexamples supply diagnostic information. We use this information to improve the strategy by iteratively training the RNN. Numerical experiments show that the proposed method elevates the state of the art in POMDP solving by up to three orders of magnitude in terms of solving times and model sizes.

3.9 Implicitly Learning to Reason in First-Order Logic

Brendan Juba (Washington University – St. Louis, US)

Joint work of Vaishak Belle, Brendan Juba
Main reference Vaishak Belle, Brendan Juba: "Implicitly learning to reason in first-order logic", pp. 3376–3386, Curran Associates, Inc., 2019.
URL http://papers.nips.cc/paper/8599-implicitly-learning-to-reason-in-first-order-logic.pdf

We consider the problem of answering queries about formulas of first-order logic based on background knowledge partially represented explicitly as other formulas, and partially represented as examples independently drawn from a fixed probability distribution. PAC semantics, introduced by Valiant, is one rigorous, general proposal for learning to reason in formal languages: although weaker than classical entailment, it allows for a powerful model theoretic framework for answering queries while requiring minimal assumptions about the form of the distribution in question. To date, however, the most significant limitation of that approach, and more generally most machine learning approaches with robustness guarantees, is that the logical language is ultimately essentially propositional, with finitely many atoms. Indeed, the theoretical findings on the learning of relational theories in such generality have been resoundingly negative. This is despite the fact that first-order logic is widely argued to be most appropriate for representing human knowledge. In this work, we present a new theoretical approach to robustly learning to reason in first-order logic, and consider universally quantified clauses over a countably infinite domain. Our results exploit symmetries exhibited by constants in the language, and generalize the notion of implicit learnability to show how queries can be computed against (implicitly) learned first-order background knowledge.

3.10 DeepProbLog: Integrating Logic, Probability and Neural Networks

Angelika Kimmig (Cardiff University, GB)

Joint work of Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, Luc De Raedt
Main reference Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, Luc De Raedt: "DeepProbLog: Neural Probabilistic Logic Programming", CoRR, Vol. abs/1907.08194, 2019.
URL https://arxiv.org/abs/1907.08194

ProbLog is a probabilistic programming language that extends the logic programming language Prolog. As other probabilistic programming and statistical relational AI techniques, it supports inference and learning. It has recently been extended to incorporate also neural networks in the framework of DeepProbLog. The resulting framework tightly integrates logic, probability and neural networks and supports both learning and reasoning and the symbolic and subsymbolic level.

3.11 Learning Description Logic Concepts: Complexity and (Un)decidability

Carsten Lutz (Universität Bremen, DE)

Joint work of Maurice Funk, Jean Christoph Jung, Carsten Lutz, Hadrien Pulcini, Frank Wolter
Main reference Maurice Funk, Jean Christoph Jung, Carsten Lutz, Hadrien Pulcini, Frank Wolter: "Learning Description Logic Concepts: When can Positive and Negative Examples be Separated?", in Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, pp. 1682–1688, ijcai.org, 2019.
URL https://doi.org/10.24963/ijcai.2019/233

Learning description logic (DL) concepts from positive and negative examples given in the form of labeled data items in a KB has received significant attention in the literature. We study the question of when a separating DL concept exists and provide useful model-theoretic characterizations as well as complexity results for the associated decision problem. For expressive DLs such as ALC and ALCQI, our characterizations show a surprising link to the evaluation of ontology-mediated conjunctive queries. We exploit this to determine the combined complexity and data complexity of separability, including a surprising undecidability result for a common DL with rather modest expressive power.

3.12 Intuitive Mathematics: Building a Proof System with Deep Reinforcement Learning

Mateusz Malinowski (Google DeepMind – London, GB)

Joint work of Alhussein Fawzi, Mateusz Malinowski, Hamza Fawzi, Omar Fawzi
Main reference Alhussein Fawzi, Mateusz Malinowski, Hamza Fawzi, Omar Fawzi: "Learning dynamic polynomial proofs", CoRR, Vol. abs/1906.01681, 2019.
URL https://arxiv.org/abs/1906.01681

Deep reinforcement learning that combines two learning paradigms is a promising method to solve complex problems that often escape the traditional formalism. With minimal domain specification and many data points, it has been shown to work effectively in the domain of complex video games. The same learning paradigm can also be used to improve the search for suitable tactics in the existing proof systems. However, I believe we can step even further and think of an end-to-end proof system. I also believe that not only theorem provers can benefit from deep reinforcement learning, but also that the former can be an excellent testbed for the latter. This talk is divided into two parts. In the first part, I will share my experience from making an algebraic proof system (I attach the corresponding paper) with deep reinforcement learning. I will mostly pay attention to 1) the reinforcement learning part, 2) incorporating inductive biases into a learned representation. The second part of my talk is more speculative. Here, I share my thoughts on building such a system that is more inspired by a human development process. The core idea relies not only on using deep reinforcement learning for more efficient search, but also to encapsulate "mathematical intuition" in the learned model.

3.13 Learning Models over Relational Databases

Dan Olteanu (University of Oxford, GB)

Joint work of Dan Olteanu, Maximilian Schleich (Oxford), Mahmoud Abo-Khamis, Ryan Curtin, Hung Q. Ngo (RelationalAI), Ben Moseley (CMU), XuanLong Nguyen (Michigan)

I will make the case for a first-principles approach to machine learning over relational databases that exploits recent development in database systems and theory. The input to learning classification and regression models is defined by feature extraction queries over relational databases. The mainstream approach to learning over relational data is to materialize the training dataset, export it out of the database, and then learn over it using statistical software packages. These three steps are expensive and unnecessary. Instead, one can cast the machine learning problem as a database problem by decomposing the learning task into a batch of aggregates over the feature extraction query and by computing this batch over the input database. Ongoing results show that the performance of this approach benefits tremendously from structural properties of the relational data and of the feature extraction query; such properties may be algebraic (semi-ring), combinatorial (hypertree width), or statistical (sampling). It also benefits from factorized query evaluation and query compilation. For a variety of models, including factorization machines, decision trees, and support vector machines, this approach may come with lower computational complexity than the materialization of the training dataset used by the mainstream approach. This translates to several orders-of-magnitude speed-up over state-of-the-art systems such as TensorFlow, R, Scikit-learn, and mlpack. While these results are promising, there is much more awaiting to be discovered.

This work is part of the FDB project.

3.14 Learning ontologies: a question-answer game

Ana Ozaki (Free University of Bozen-Bolzano, IT)

Joint work of Ricardo Duarte, Boris Konev, Carsten Lutz, Ana Ozaki, Frank Wolter

Ontologies have been applied to integrate and abstract information from multiple data sources; to describe knowledge in various domains, in particular, those related life sciences; among others. Building an ontology often requires the interaction between experts in a domain of interest and experts in modelling ontologies, called ontology engineers. We treat the problem of building an ontology as a learning problem. An ontology engineer, playing the role of the learner, attempts to build an ontology that reflects the knowledge of a domain expert (the teacher) by posing questions. This setting can be seen as an instance of Angluin's exact learning model with membership and equivalence queries. We investigate polynomial learnability for different ontology languages within this learning model and show non-polynomial learnability for ontologies formulated in the ontology language EL, and polynomial learnability for fragments of this language. We also present an implementation of an (exponential) algorithm for learning EL ontologies.

The talk will be primarily based upon the work in [1, 2].

References

1 Boris Konev, Carsten Lutz, Ana Ozaki, Frank Wolter, "Exact Learning of Lightweight Description Logic Ontologies", Journal of Machine Learning Research, Vol. 18(201), Pages 1–63, 2018.

2 Mario Ricardo Cruz Duarte, Boris Konev, Ana Ozaki, "ExactLearner: A Tool for Exact Learning of EL Ontologies", Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR, Pages 409–414, 2018.

3.15 Learning Logics, Program Synthesis, and Neural Nets

Madhusudan Parthasarathy (University of Illinois – Urbana-Champaign, US)

This talk will survey three fields that have slightly different emphasis: learning logics (learning formulas from data), program synthesis (especially using learning), and neural nets (for recognizing patterns). I will try to explore these areas, their applications we have pursued (synthesizing programs, synthesizing inductive, mining specifications), and new synergies that suggest a new kind of intelligence that combines neural inductive learning and symbolic learning of interpretable concepts that can be used for reasoning with applications to a more general artificial intelligence.

3.16 Entity Resolution: A Case for Logic and Learning

Lucian Popa (IBM Almaden Center – San Jose, US)

Joint work of Douglas Burdick, Ronald Fagin, Sairam Gurajada, Jungo Kasai, Phokion Kolaitis, Yunyao Li, Lucian Popa, Kun Qian, Prithviraj Sen, Wang-Chiew Tan
Main reference Kun Qian, Lucian Popa, Prithviraj Sen: "Active Learning for Large-Scale Entity Resolution", in Proc. of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 – 10, 2017, pp. 1379–1388, ACM, 2017.
URL https://doi.org/10.1145/3132847.3132949

Entity resolution is a key form of reasoning over data that allows to establish explicit connections among entities across diverse datasets. In this talk, I will make the case that building good abstractions and tools for entity resolution requires a combination of logic-based methods and machine learning techniques. I will briefly describe a declarative approach that uses constraints and provides a logical foundation for reasoning about entity resolution specifications and their expressive power. This also forms the theoretical underpinning for a concrete high-level language that is used in production by IBM. I will then talk about learning techniques to facilitate the generation of good entity resolution programs using the logic-based language as the target. Of particular importance are active learning techniques where the machine and the human-expert cooperate in order to reach high-accuracy entity resolution algorithms in concrete application scenarios.

3.17 Synthesizing Datalog Programs using Numerical Relaxation

Xujie Si (University of Pennsylvania – Philadelphia, US)

Joint work of Xujie Si, Mukund Raghothaman, Kihong Heo, Mayur Naik
Main reference Xujie Si, Mukund Raghothaman, Kihong Heo, Mayur Naik: "Synthesizing Datalog Programs using Numerical Relaxation", in Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, pp. 6117–6124, ijcai.org, 2019.
URL https://doi.org/10.24963/ijcai.2019/847

The problem of learning logical rules from examples arises in diverse fields, including program synthesis, logic programming, and machine learning. Existing approaches either involve solving computationally difficult combinatorial problems or performing parameter estimation in complex statistical models. In this paper, we present DIFFLOG, a technique to extend the logic programming language Datalog to the continuous setting. By attaching real-valued weights to individual rules of a Datalog program, we naturally associate numerical values with individual conclusions of the program. Analogous to the strategy of numerical relaxation in optimization problems, we can now first determine the rule weights which cause the best agreement between the training labels and the induced values of output tuples, and subsequently recover the classical discrete-valued target program from the continuous optimum. We evaluate DIFFLOG on a suite of 34 benchmark problems from recent literature in knowledge discovery, formal verification, and database query-by-example, and demonstrate significant improvements in learning complex programs with recursive rules, invented predicates, and relations of arbitrary arity.

3.18 Information Theory and Data Management

Dan Suciu (University of Washington – Seattle, US)

Joint work of Dan Suciu, Batya Kenig
Main reference Batya Kenig, Dan Suciu: "Integrity Constraints Revisited: From Exact to Approximate Implication", CoRR, Vol. abs/1812.09987, 2018.
URL https://arxiv.org/abs/1812.09987

I will describe three applications of Information Theory to Data Management: upper bounds on query size, approximate constraints, and containment of queries with bag semantics.

3.19 Higher Order Theorem Proving by Deep Learning

Christian Szegedy (Google Inc. – Mountain View, US)

Joint work of Aditya Paliwal, Christian Szegedy, Dennis Lee, Kshitij Bansal, Markus Rabe, Sarah Loos, Stewart Wilcox, Viktor Toman, Kshitij Bansal
Main reference Kshitij Bansal, Sarah M. Loos, Markus N. Rabe, Christian Szegedy, Stewart Wilcox: "HOList: An Environment for Machine Learning of Higher Order Logic Theorem Proving", in Proc. of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, Proceedings of Machine Learning Research, Vol. 97, pp. 454–463, PMLR, 2019.
URL http://proceedings.mlr.press/v97/bansal19a.html

I give an overview of the HOList benchmark and the DeepHOL system for fully automated theorem proving for higher order logic in large theories using a tactic based prover trained by deep reinforcement learning. I will discuss recent results on exploration based strategies for theorem proving. This avoids the necessity of imitation learning on human prooflogs. It is also demonstrates how the choice of suitable deep learning model architecture affects the overall proving performance significantly.

3.20 Machine Learning and Knowledge Graphs

Balder Ten Cate (Google Inc. – Mountain View, US)

In the light of recent developments in (deep) Machine Learning (ML) involving attention mechanisms, and pretraining and finetuning of models, there is a question of whether knowledge bases can add any value, since it has already been shown that these ML models can already learn to answer queries directly from documents. In this talk, I will discuss several advantages that Knowledge Graphs (KG) still offer, such as interoperability, stability over time, and controllability. I will then turn to the question of how ML and KGs can work together, and how ML models can learn to use the data present in a KG. I discuss some approaches to solving these problems and present a high level overview of different possible interfaces between ML and KG.

3.21 Combining Learning and Reasoning over Large Formal Math Corpora

Josef Urban (Czech Technical University – Prague, CZ)

The talk will start with a brief motivation for building strong AI for math and science via combining learning and reasoning over large formal mathematical corpora created with proof assistants such as Mizar, Isabelle, HOL and Coq. I will then describe several tasks in this area such as learning of premise selection over large libraries, learning to guide saturation-style and tableau-style automated theorem provers (ATPs), learning to guide tactical interactive theorem provers, learning of theorem proving strategies, conjecturing, etc. I will also mention various feedback loops between proving and learning in some of these settings, and show some of our autoformalization experiments.

3.22 Statistical Relational Learning

Guy Van den Broeck (UCLA, US)

This talk discusses the role of logical reasoning in statistical machine learning. While their unification has been a long-standing and crucial open problem, automated reasoning and machine learning are still disparate fields within artificial intelligence. I will describe recent progress towards their synthesis in several facets. I start with a very practical question:

how can we enforce logical constraints on the output of deep neural networks to incorporate symbolic knowledge? Second, I explain how circuits developed for tractable logical reasoning can be turned into statistical models. When brought to bear on a variety of machine learning tasks, including discrete density estimation and simple image classification, these probabilistic and logistic circuits yield state-of-the-art results. Finally I give a brief overview of statistical relational learning.

3.23 Towards Finding Longer Proofs

Zsolt Zombori (Alfréd Rényi Institute of Mathematics – Budapest, HU)

Joint work of Zsolt Zombori, Adrián Csiszárik, Henryk Michalewski, Cezary Kaliszyk, Josef Urban
Main reference Zsolt Zombori, Adrián Csiszárik, Henryk Michalewski, Cezary Kaliszyk, Josef Urban: "Towards Finding Longer Proofs", CoRR, Vol. abs/1905.13100, 2019.
URL https://arxiv.org/abs/1905.13100

I present a reinforcement learning (RL) based guidance system for automated theorem proving geared towards Finding Longer Proofs (FLoP). FLoP focuses on generalizing from short proofs to longer ones of similar structure. To achieve that, FLoP uses state-of-the-art RL approaches that were previously not applied in theorem proving. In particular, we show that curriculum learning significantly outperforms previous learning-based proof guidance on a synthetic dataset of increasingly difficult arithmetic problems.

4 Breakout Sessions

4.1 Differentiable FOL Learning from Structures

Adithya Murali (University of Illinois – Urbana-Champaign, US)

Joint work of All participants

Logical structures possess many advantages: they are highly interpretable (and can therefore be inspected or studied in detail), highly compositional, and have many data-efficient learning algorithms. Most importantly, in the context of many AI problems such as analogy reasoning, policy learning or even simple classification, logic offers many excellent modelling choices that can abstract higher-order patterns over primitives. These primitives could correspond to complex non-logical entities such as visual inputs [1] or other signals. While all of these characteristics are well-known and were used in AI systems many years ago, the common criticism is that they are extremely intolerant to noise and, traditionally, offer no way of expressing something like *approximate satisfiability* with respect to a concept. As rightly observed by the authors in [2], they also typically cannot naturally (directly) handle ambiguous non-symbolic data such as raw pixel inputs.

In the last few years, differentiable programming has emerged as a framework for program induction. In this setup, the class of programs is defined by low-level differentiable function families that can be combined by simple higher-level programmatic combinators, and the form of the program is learnt using gradient descent. These programmatic combinators could be higher-order functions like *map* or *fold* [3], but could also be tape head movement or memory updates as in the work on Neural Turing Machines and Differentiable Neural Computers [4, 5, 6]. The learnt concept usually possesses some programmatic structure that can indicate the logic behind the learnt solution, but the crucial inductive generalisations and patterns are learnt from data and are contained in the weights of a neural network. There is therefore a natural interest to apply this philosophy to learning logics, and the breakout session on differentiable FOL (First-Order Logic) learning from structures was centered around this interest. The research question was whether first-order logic formulae could be learnt – as classifiers discriminating between a few first-order structures classified as positive or negative examples – in a *differentiable* manner.

The session began with an introduction to the context of the problem and the desired goals as described above. The first point of discussion was the possibility of being able to embed formulae into a real vector space. The argument against this was that meaningful embeddings (for example, those that mapped semantically equivalent formulae to the same vector) seemed difficult to obtain in a general way. The suggestion was to then obtain these embeddings using a neural network, similar to word embeddings [7, 8] and code embeddings [9]. This was deemed an unsatisfactory solution towards the desired goals as it was unclear how one would obtain a classifier at the end using this method.

Two contrary positions were offered. The first one was that one could embed the structures instead into a (perhaps high-dimensional) space and search for particular classes of formulae that would correspond to tractable structures in that space. However, it was unclear how the learning would be done in a differentiable manner in that setting. The second suggestion received more interest and essentially posited that the semantics of the logic itself could be *lifted* to a continuous or differentiable setting. This would give us the desired 'approximate satisfiability' and could be used to define a metric representing the ability of the formula to discriminate between the given examples, which could then be used to learn the formula that minimises that metric using gradient descent. There is work in this direction [2], including some that were part of talks at the seminar [10, 11]. Some merits and demerits of these works were discussed.

The next question that was addressed was where the data would come from to train such a system, and several suggestions were offered:

- Databases (where the task would be to learn SQL queries).
- Knowledge Graphs. The criticism was that such graphs are large and few, and would therefore not be a source for enough examples. It was not clear whether one could use subgraphs of these graphs to generate more data.
- Randomly generated structures and discriminators. However, it was not clear whether this would generate sufficiently many examples since First-Order Logic follows a Zero-One law and it could happen that either the generated formula would hold on either structure or neither one.
- Datasets like the Visual Genome image dataset [12]. This suggestion was well-received since it would provide enough examples and was an interesting application domain.

The final segment of the breakout session was on suitable problems and ambitious research questions towards differentiably learning FOL formulae. One example provided was that of problems in the verification domain such as invariant synthesis or precondition generation, since the problem statements already require the output to be formulae. Another example was that of generalisation to unseen combinations of properties [13], where a logical representation would naturally be better at expressing such combinations and differentiable learning would remove the need to craft the individual properties by hand. The session concluded with the following question: the methods and works indicated so far do not handle quantifiers; how would one differentiably learn FOL formulae that have quantifiers?

References

1 Adithya Murali, P. Madhusudan, "Augmenting Neural Nets with Symbolic Synthesis: Applications to Few-Shot Learning", CoRR, abs/1907.05878, July 2019.

2 Richard Evans, Edward Grefenstette, "Learning explanatory rules from noisy data", Journal of Artificial Intelligence Research archive, Volume 61 Issue 1, Pages 1-64, January 2018.

3 Lazar Valkov, Dipak Chaudhari, Akash Srivastava, Charles A. Sutton, Swarat Chaudhuri, "Houdini: Lifelong Learning as Program Synthesis", Neural Information Processing Systems (NeurIPS), 2018.

4 Alex Graves, Greg Wayne, Ivo Danihelka, "Neural Turing Machines", CoRR, abs/1410.5401, December 2014.

5 Mark Collier, Joeran Beel, "Implementing Neural Turing Machines", 27th International Conference on Artificial Neural Networks (ICANN), 2018.

6 Alex Graves et al., "Hybrid computing using a neural network with dynamic external memory", Nature, Vol. 538, Pages 471–476, October 2016.

7 Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient estimation of word representations in vector space", CoRR, abs/1301.3781, 2013.

8 Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, Jeffrey Dean, "Distributed representations of words and phrases and their compositionality", Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, pages 3111–3119, December 2013.

9 Uri Alon, Meital Zilberstein, Omer Levy, Eran Yahav, "code2vec: Learning Distributed Representations of Code", Proceedings of the ACM on Programming Languages, Volume 3 Issue POPL, January 2019.

10 Xujie Si, Mukund Raghothaman, Kihong Heo, Mayur Naik, "Synthesizing Datalog Programs Using Numerical Relaxation", Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19), Pages 6117-6124, August 2019.

11 Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, Luc De Raedt, "DeepProbLog: Neural Probabilistic Logic Programming", NeurIPS 2018.

12 Ranjay Krishna et al., "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations", International Journal of Computer Vision, Volume 123 Issue 1, Pages 32-73, May 2017.

13 Junhyuk Oh, Satinder Singh, Honglak Lee, Pushmeet Kohli, "Zero-Shot Task Generalization with Multi-Task Deep Reinforcement Learning", International Conference on Machine Learning (ICML), 2017.

4.2 Explainable AI

Adithya Murali (University of Illinois – Urbana-Champaign, US)

Joint work of All participants

Supervised and semi-supervised deep learning systems such as AlphaGo Zero [1] have surpassed expectations in many fields, including gameplay, self-driving cars, and medical diagnosis. This has prompted the concerns of individuals and governments alike on the decisions made by so-called "black-box" models, or more generally *Explainable AI* (XAI). There have been many attempts through the years [2, 3, 4] to define and characterise in various application domains the concept of an *explanation*, which was the prime focus of the breakout session on XAI.

The primary and perhaps simplistic suggestion was that explanations could be formulae, programs, or constraints – or more generally symbolic objects that are highly interpretable. This of course presents numerous problems, including the fact that such objects, while highly compositional, do not usually offer forms of "approximate satisfaction" with respect to a concept. A different suggestion was to relax the symbolic requirement by instead using objects that are sub-symbolic but still possess some compositional structure. An example of this would be high-level programmatic structures over non-symbolic primitives as in the spirit of works such as [5, 6]. A contrary suggestion posited that explanations could be more complex objects such as dialogue, using even a "black-box" internal representation to continuously clarify and detail the output in the context of questions asked by an agent seeking explanation. However, the argument was made that this would not be an XAI system, but rather an AI system that produces explanations. This distinction indicated the possibility of XAI being an AI-complete problem.

The discussion then turned to some alternative strategies to further the task of exploring XAI. Since attempts to find an all-encompassing definition seemed to fail or were too trivial, one suggestion was to instead attempt a characterisation of explanations by finding a desired set of properties or axioms. Combinations of these axioms would help determine structures that could act as explanations for various applications. One could then use this setup to abstractly study the properties of such "explanation structures" and prove theorems about them. This is in the spirit of works such as [7], which defined a family of fairness measures using axioms.

The last segment of the session was on possible applications or problems where explanations might be necessary. One such example that was suggested was that of chess engines, where one could require not merely gameplay but the extraction of concrete strategies spanning across many moves or explanations for positions that the engines would rate as advantageous for one player. A more interesting suggestion pertained to games such as Angry Birds™ that are played in rounds, where the player would modify their strategy slightly based on the successes or failures of their attempts in earlier rounds. The discussion concluded with a reading of the main points made during the session.

References

1 David Silver et al., "Mastering the game of Go without human knowledge", Nature, Vol. 550, October 2017.

2 Michael van Lent, William Fisher, Michael Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior", Proceedings of the National Conference on Artificial Intelligence, pp. 900-907, July 2004.

3 Mustafa Bilgic, Raymond J. Mooney, "Explaining Recommendations: Satisfaction vs. Promotion", Proceedings of Beyond Personalization 2005, the Workshop on the Next Stage of Recommender Systems Research (IUI2005), January 2005.

4 Nazneen Fatema Rajani, Raymond J. Mooney, "Ensembling Visual Explanations for VQA", Proceedings of the NeurIPS 2017 workshop on Visually-Grounded Interaction and Language (ViGIL), December 2017.

5 Brenden M. Lake, Ruslan Salakhutdinov, Joshua B. Tenenbaum, "Human-level concept learning through probabilistic program induction", Science, Vol. 350, Issue 6266, pp. 1332-1338, December 2015.

6 Adithya Murali, P. Madhusudan, “Augmenting Neural Nets with Symbolic Synthesis: Applications to Few-Shot Learning”, CoRR, abs/1907.05878, July 2019.

7 Tian Lan, David Kao, Mung Chiang, Ashutosh Sabharwal, “An axiomatic theory of fairness in network resource allocation”, Proceedings of the 29th conference on Information communications (INFOCOM’10), pages 1343-1351, March 2010.

4.3 Injecting Symbolic Knowledge/Constraints into Neural Networks

Adithya Murali (University of Illinois – Urbana-Champaign, US)

Joint work of All participants

While connectionist and computationalist methods are often spoken about as being at odds with each other, there have always been efforts to find techniques that reconcile them in some manner. One of the strategies proposed often is to *constrain* the training or output of neural networks using some logical theories or constraints. This was the position that was taken by the participants of the breakout session on injecting symbolic knowledge/constraints into neural networks.

The discussion began with the mention of early works such as Knowledge-Based Artificial Neural Networks (KBANN) [1], which incorporated certain “domain theories” represented in propositional logic into neural networks, building *hybrid learning systems* that used two kinds of information sources: structured knowledge in the form of logic as well as a set of classified examples. The authors of that work showed that these models could learn to generalise to unseen examples better than those using only one kind of information source. Recent works such as the work on Adviceptron [2] or Knowledge-Based Probabilistic Logic Learning [3] have developed on this philosophy by relaxing the hard symbolic constraints to soft constraints, or generalising it to noisy domains.

Then, the desired goals of such an injection of symbolic knowledge into neural networks were discussed. There were at least three clear goals:

1. To speed up or robustify the learning process of neural networks by using symbolic knowledge, similar to the goals of works in [2, 4].
2. To obtain a model at the end of the learning/training process that either incorporates, remembers, or in some way satisfies the given constraints. The choice of defining the appropriate constraints could, however, be domain-specific. For example, the constraints could enforce a particular hypothesis class or representation for the output (say, programs of some bounded measure) as in the work of [5]. They could also be soft constraints that essentially encode an objective function that measures the learnt model’s ability to satisfy the real (hard) constraints as in the work of [6].
3. To attempt at building mechanisms of interaction between neural networks and symbolic knowledge modules, perhaps by providing the neural network access to symbolic reasoning engines (similar to the work in [7]).

The rest of the discussion essentially focused on the appropriate language or *medium* to express constraints. There were a few different positions expressed upon this subject. Experts from the Machine Learning community were divided on the position of injecting knowledge using the loss function of the model training phase. Some experts argued that finding the right loss function that approximates (in limit) the desired constraints would be enough.

While this is fairly standard practice, careful analysis of the loss functions using theorems about their properties does not appear to be standard practice. It was also argued that such a methodology was currently still a creativity-based approach rather than a systematic one. There are certainly exceptions as in the work of [6] (which also formed a central theme of one of the seminar talks) where the authors defined a 'semantic loss' that lifts logical constraints with Boolean satisfiability into a continuous form where the constraints could be satisfied fairly well or poorly. However, a crucial argument of those experts against entirely depending on loss functions was that if the data had different patterns or correlations, or in some sense had opposing conclusions to that of the loss function, it would be quite impossible for the training to successfully bridge the two. In general it was possible that the training would jump back forth between the two possibilities and not really converge.

The discussion then turned to a few orthogonal approaches, suggesting that constraints could be richer than simple input/output examples or logically expressed properties. For instance, they could express templates for neural networks as in the work of [8, 9]. Yet another suggestion was that one could, instead of targeting an injunction of knowledge, target extraction of knowledge by coming up with a general formal framework for verifying machine learning models. There is work in this direction [10], but it appeared that most participants disagreed with extraction as against injunction for the purposes of the goals illustrated above.

The session concluded with some thoughts about situations or challenge problems where injecting knowledge would be essential.

References

1 Geoffrey G. Towell, Jude W. Shavlik, "Knowledge-based artificial neural networks", Artificial Intelligence archive, Volume 70 Issue 1-2, Pages 119 – 165, October 1994.

2 G. Kunapuli, K. P. Bennett, R. Maclin, J. W. Shavlik, "The Adviceptron: Giving Advice To The Perceptron", Twentieth Conference on Artificial Neural Networks In Engineering (ANNIE'10), November 2010.

3 Phillip Odom, Tushar Khot, Reid Porter, Sriraam Natarajan, "Knowledge-based probabilistic logic learning", Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15), Pages 3564-3570, January 2015.

4 Jialin Wu, Raymond J. Mooney, "Self-Critical Reasoning for Robust Visual Question Answering", Proceedings of Neural Information Processing Systems (NeurIPS), December 2019.

5 Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, Swarat Chaudhuri, "Programmatically Interpretable Reinforcement Learning", Proceedings of the 35th International Conference on Machine Learning (ICML), 2018.

6 Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, Guy Van den Broeck, "A Semantic Loss Function for Deep Learning with Symbolic Knowledge", Proceedings of the 35th International Conference on Machine Learning (ICML), 2018.

7 Kevin Ellis, Daniel Ritchie, Armando Solar-Lezama, Joshua B. Tenenbaum, "Learning to Infer Graphics Programs from Hand-Drawn Images", Advances in Neural Information Processing Systems 31 (NeurIPS 2018), 2018.

8 Gustav Sourek, Vojtech Aschenbrenner, Filip Zelezny, Ondrej Kuzelka, "Lifted Relational Neural Networks", COCO'15 Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches – Volume 1583, Pages 52-60, December 2015.

9 Gustav Sourek, Vojtech Aschenbrenner, Filip Zelezny, Steven Schockaert, Ondrej Kuzelka, "Lifted relational neural networks: efficient learning of latent relational structures", Journal of Artificial Intelligence Research archive, Volume 62 Issue 1, Pages 69-100, May 2018.

10 Guy Katz et al., "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks", Computer Aided Verification – 29th International Conference (CAV 2017) July 2017.

Participants

- Isolde Adler
University of Leeds, GB
- Molham Aref
relational AI – Berkeley, US
- Vaishak Belle
University of Edinburgh, GB
- Michael Benedikt
University of Oxford, GB
- Ismail Ilkan Ceylan
University of Oxford, GB
- Victor Dalmau
UPF – Barcelona, ES
- Luc De Raedt
KU Leuven, BE
- Dana Fisman
Ben Gurion University –
Beer Sheva, IL
- James Freitag
University of Illinois –
Chicago, US
- Ivan Gavran
MPI-SWS – Kaiserslautern, DE
- Martin Grohe
RWTH Aachen, DE
- Barbara Hammer
Universität Bielefeld, DE
- Daniel Huang
University of California –
Berkeley, US
- Nils Jansen
Radboud University
Nijmegen, NL
- Brendan Juba
Washington University –
St. Louis, US
- Kristian Kersting
TU Darmstadt, DE
- Sandra Kiefer
RWTH Aachen, DE
- Angelika Kimmig
Cardiff University, GB
- Phokion G. Kolaitis
University of California – Santa
Cruz & IBM Almaden Research
Center – San Jose, US
- Egor Kostylev
University of Oxford, GB
- Paul Krogmeier
University of Illinois –
Urbana Champaign, US
- Luis C. Lamb
Federal University of
Rio Grande do Sul, BR
- Carsten Lutz
Universität Bremen, DE
- Mateusz Malinowski
Google DeepMind – London, GB
- Henryk Michalewski
University of Warsaw, PL
- Adithya Murali
University of Illinois –
Urbana-Champaign, US
- Sriraam Natarajan
University of Texas – Dallas, US
- Daniel Neider
MPI-SWS – Kaiserslautern, DE
- Dan Olteanu
University of Oxford, GB
- Ana Ozaki
Free University of Bozen-Bolzano, IT
- Madhusudan Parthasarathy
University of Illinois –
Urbana-Champaign, US
- Lucian Popa
IBM Almaden Center –
San Jose, US
- Martin Ritzert
RWTH Aachen, DE
- Xujie Si
University of Pennsylvania –
Philadelphia, US
- Dan Suciu
University of Washington –
Seattle, US
- Christian Szegedy
Google Inc. –
Mountain View, US
- Balder Ten Cate
Google Inc. –
Mountain View, US
- Josef Urban
Czech Technical University –
Prague, CZ
- Steffen van Bergerem
RWTH Aachen, DE
- Guy Van den Broeck
UCLA, US
- Zsolt Zombori
Alfréd Rényi Institute of
Mathematics – Budapest, HU

Report from Dagstuhl Seminar 19371

Deduction Beyond Satisfiability

Edited by

Carsten Fuhs[1], Philipp Rümmer[2], Renate Schmidt[3], and Cesare Tinelli[4]

1 **Birkbeck, University of London, GB,** carsten@dcs.bbk.ac.uk
2 **Uppsala University, SE,** philipp.ruemmer@it.uu.se
3 **University of Manchester, GB,** renate.schmidt@manchester.ac.uk
4 **University of Iowa – Iowa City, US,** cesare-tinelli@uiowa.edu

Abstract

Research in automated deduction is traditionally focused on the problem of determining the satisfiability of formulas or, more generally, on solving logical problems with yes/no answers. Thanks to recent advances that have dramatically increased the power of automated deduction tools, there is now a growing interest in extending deduction techniques to attack logical problems with more complex answers. These include both problems with a long history, such as quantifier elimination, which are now being revisited in light of the new methods, as well as newer problems such as minimal unsatisfiable cores computation, model counting for propositional or first-order formulas, Boolean or SMT constraint optimization, generation of interpolants, abductive reasoning, and syntax-guided synthesis. Such problems arise in a variety of applications including the analysis of probabilistic systems (where properties like safety or liveness can be established only probabilistically), network verification (with relies on model counting), the computation of tight complexity bounds for programs, program synthesis, model checking (where interpolation or abductive reasoning can be used to achieve scale), and ontology-based information processing. The seminar brought together researchers and practitioners from many of the often disjoint subcommunities interested in the problems above. The unifying theme of the seminar was how to harness and extend the power of automated deduction methods to solve problems with more complex answers than binary ones.

Seminar September 8–13, 2019 – http://www.dagstuhl.de/19371

2012 ACM Subject Classification Theory of computation → Automated reasoning, Computing methodologies → Knowledge representation and reasoning, Theory of computation → Logic and verification

Keywords and phrases abduction, automated deduction, interpolation, quantifier elimination, synthesis

Digital Object Identifier 10.4230/DagRep.9.9.23

Edited in cooperation with Warren Del-Pinto

Deduction Beyond Satisfiability, *Dagstuhl Reports*, Vol. 9, Issue 9, pp. 23–44
Editors: Carsten Fuhs, Philipp Rümmer, Renate Schmidt, and Cesare Tinelli
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Carsten Fuhs
Philipp Rümmer
Renate Schmidt
Cesare Tinelli

This report contains the program and outcomes of Dagstuhl Seminar 19371 on "Deduction Beyond Satisfiability" held at Schloss Dagstuhl, Leibniz Center for Informatics, during September 10–15, 2017. It was the thirteenth in a series of Dagstuhl Deduction seminars held biennially since 1993.

Research in automated deduction has traditionally focused on solving decision problems, which are problems with a binary answer. Prominent examples include proving the unsatisfiability of a formula, proving that a formula follows logically from others, checking the consistency of an ontology, proving safety or termination properties of programs, and so on. However, automated deduction methods and tools are increasingly being used to address problems with more complex answers, for instance to generate programs from formal specifications, compute complexity bounds, or find optimal solutions to constraint satisfaction problems.

In some cases, the required extended functionality (e.g., to identify unsatisfiable cores) can be provided relatively easily from current deduction procedures. In other cases (e.g., for Craig interpolation, or to find optimal solutions of constraints), elaborate extensions of these procedures are needed. Sometimes, altogether different methods have to be devised (e.g., to count the number of models of a formula, compute the set of all consequences of an ontology, identify missing information in a knowledge base, transform and mine proofs, or analyze probabilistic systems). In all cases, the step from yes/no answers to such extended queries and complex output drastically widens the application domain of deductive machinery. This is proving to be a key enabler in a variety of areas such as formal methods (for software/hardware development) and knowledge representation and reasoning.

While promising progress has been made, many challenges remain. Extending automated deduction methods and tools to support these new functionalities is often intrinsically difficult, and challenging both in theory and implementation. The scarcity of interactions between the involved sub-communities represents another substantial impediment to further advances, which is unfortunate because these sub-communities often face similar problems and so could greatly benefit from the cross-fertilization of ideas and approaches. An additional challenge is the lack of common standards for interfacing tools supporting the extended queries. Developing common formalisms, possibly as extensions of current standard languages, could be as transformative to the field as the introduction of standards such as TPTP and SMT-LIB has been in the past.

This Dagstuhl seminar brought together researchers working on deduction methods and tools that go beyond satisfiability and other traditional decision problems; specialists that work on advanced techniques in deduction and automated reasoning such as model counting, quantifier elimination, interpolation, abduction, or optimization; and consumers of deduction technology who need answers to more complex queries than just yes/no questions.

The unifying theme of the seminar was how to harness and extend the power of automated deduction methods to solve a variety of non-decision problems with useful applications. Research questions addressed at the seminar were the following:

- What kind of information should be passed to a "beyond satisfiability" deduction tool, and what information should be returned to the user? The goal should be to enhance the understanding of related concepts in different subfields and applications, and to converge towards common formalisms.
- How can current ideas, results and systems in one sub-community of researchers and practitioners benefit the needs of other communities?
- What are outstanding challenges in using and building deduction tools to attack logical problems with complex answers?

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 Safe Decomposition of Startup Requirements: Verification and Synthesis

Alberto Griggio (Bruno Kessler Foundation – Trento, IT)

Joint work of Alberto Griggio, Alessandro Cimatti, Luca Geatti, Stefano Tonett

The initialization of complex cyber-physical systems often requires the interaction of various components that must start up with strict timing requirements on the provision of signals (power, refrigeration, light, etc.). In order to safely allow an independent development of components, it is necessary to ensure a safe decomposition, i.e. the specification of local timing requirements that prevent later integration errors due to the dependencies.

We propose a high-level formalism to model local timing requirements and dependencies. We consider the problem of checking the consistency (existence of an execution satisfying the requirements) and compatibility (absence of an execution that reach an integration error) of the local requirements, and the problem of synthesizing a region of timing constraints that represent all possible correct refinements of the original specification. We show how the problems can be naturally translated into a reachability and synthesis problem for timed automata with shared variables. Exploiting the linear structure of the requirements, we propose an encoding of the problem into SMT. We evaluate the SMT-based approach using Mathsat and show how it scales better compared to the automata-based approach using Uppaal and nuXmv.

3.2 Proof Checking in Zipperposition

Alexander Bentkamp (Free University Amsterdam, NL)

Joint work of Alexander Bentkamp, Simon Cruanes

Beyond a simple yes/no answer, automated theorem provers can typically generate proof output. To detect soundness bugs and to ensure better guarantees on the correctness of these proofs, we developed a proof checker for Zipperposition, a prover for polymorphic higher-order logic. The checker reduces the proof into a series of ground problems, which can be decidably and efficiently checked by a simple SMT prover.

3.3 Guiding High-Performance SAT Solvers with Unsat-Core Predictions

Nikolaj S. Bjørner (Microsoft Research – Redmond, US)

Joint work of Nikolaj S. Bjørner, Daniel Selsam
Main reference Daniel Selsam, Nikolaj Bjørner: "Guiding High-Performance SAT Solvers with Unsat-Core Predictions", in Proc. of the Theory and Applications of Satisfiability Testing – SAT 2019 – 22nd International Conference, SAT 2019, Lisbon, Portugal, July 9-12, 2019, Proceedings, Lecture Notes in Computer Science, Vol. 11628, pp. 336–353, Springer, 2019.
URL http://dx.doi.org/10.1007/978-3-030-24258-9_24

The NeuroSAT neural network architecture was introduced by Selsam et al for predicting properties of propositional formulae. When trained to predict the satisfiability of toy problems, it was shown to find solutions and unsatisfiable cores on its own. However, the authors saw "no obvious path" to using the architecture to improve the state-of-the-art. In this work, we train a simplified NeuroSAT architecture to directly predict the unsatisfiable cores of real problems. We modify several state-of-the-art SAT solvers to periodically replace their variable activity scores with NeuroSAT's prediction of how likely the variables are to appear in an unsatisfiable core. The modifications led to speedups of 10 percent on unseen and diverse problems and 20 percent on problems form a scheduling domain. Our results demonstrate that NeuroSAT can provide effective guidance to high-performance SAT solvers on real problems. The talk introduces the techniques for representing CNF problems as graphical neural networks, our choice of training approach and some caveats in whether the measured improvements can be obtained in alternative ways. NeuroCore was developed in collaboration with Daniel Selsam and develops on the NeuroSAT architecture also developed by Selsam and collaborators.

References

1 Daniel Selsam, Matthew Lamm, Benedikt Bünz, Percy Liang, Leonardo de Moura, David L. Dill: Learning a SAT Solver from Single-Bit Supervision. ICLR 2019.

2 Daniel Selsam, Nikolaj Bjørner: Guiding High-Performance SAT Solvers with Unsat-Core Predictions. SAT 2019: 336-353

3.4 Proof reconstruction in conflict-driven satisfiability

Maria Paola Bonacina (Università degli Studi di Verona, IT)

Joint work of Maria Paola Bonacina, Stéphane Graham-Lengrand, Natarajan Shankar
Main reference Maria Paola Bonacina, Stéphane Graham-Lengrand, Natarajan Shankar: "Proofs in conflict-driven theory combination", in Proc. of the 7th ACM SIGPLAN International Conference on Certified Programs and Proofs, CPP 2018, Los Angeles, CA, USA, January 8-9, 2018, pp. 186–200, ACM, 2018.
URL http://dx.doi.org/10.1145/3167096

Proofs of unsatisfiability, or, equivalently, validity, are an important output of automated reasoning methods, and their transformation, exchange, and standardization is a key factor for the interoperability of different automated reasoning systems. In theorem proving proof reconstruction is the task of extracting a proof from the final state of a derivation after generating the empty clause, and for several theorem-proving methods it is a standard, though nontrivial, task. In SAT solving the conflict-driven clause learning procedure (CDCL) generates proofs by resolution: while this is true in principle, in practice, SAT-solver proofs are so huge that their definition, generation, and manipulation is an active research topic. In SMT solving, which represents a middle ground between first-order theorem proving and SAT solving, proof generation is also crucial, and while it receives increasing attention, the field does not seem to have a standard output format for proofs. This talk aims at contributing to this discussion by presenting approaches to proof reconstruction in CDSAT (Conflict-Driven SATisfiability), the paradigm for satisfiability modulo theories and assignments (SMT/SMA) developed by the author with Stéphane Graham-Lengrand and Natarajan Shankar.

3.5 Tractable QBF and model counting via Knowledge Compilation

Florent Capelli (INRIA Lille, FR)

Joint work of Florent Capelli, Simone Bova, Stefan Mengel, Friedrich Slivovsky
Main reference Florent Capelli, Stefan Mengel: "Tractable QBF by Knowledge Compilation", in Proc. of the 36th International Symposium on Theoretical Aspects of Computer Science, STACS 2019, March 13-16, 2019, Berlin, Germany, LIPIcs, Vol. 126, pp. 18:1–18:16, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2019.
URL http://dx.doi.org/10.4230/LIPIcs.STACS.2019.18

We show how knowledge compilation can be used as a tool for solving QBF and more. More precisely, we show that one can apply quantification on certain data structures used in knowledge compilation which in combination with the fact that restricted classes of CNF-formulas can be compiled into these data structures can be used to show fixed-parameter tractable results for QBF. In particular, we rediscover a result by Hubie Chen (ECAI 04) on FPT-tractability of QBF on bounded treewidth CNF and generalise it to aggregation problems such as counting or enumerating the models of the input quantified CNF.

This talk will review joint work with Simone Bova, Stefan Mengel and Friedrich Slivovsky.

3.6 Forgetting-Based Abductive Reasoning and Inductive Learning in Ontologies

Warren Del-Pinto (University of Manchester, GB)

Joint work of Warren Del-Pinto, Renate A. Schmidt
Main reference Warren Del-Pinto, Renate A. Schmidt: "ABox Abduction via Forgetting in ALC", in Proc. of the The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 – February 1, 2019, pp. 2768–2775, AAAI Press, 2019.
URL http://dx.doi.org/10.1609/aaai.v33i01.33012768

We present an approach to performing abductive reasoning in large description logic (DL) ontologies. The approach is based on the use of forgetting. Characteristics of the hypotheses obtained and important considerations such as redundancy elimination will be discussed alongside experimental results.

We also discuss the potential interaction between this form of abductive reasoning and inductive learning in DL ontologies. In particular: how do the characteristics of these hypotheses lend themselves to several learning problems in this setting?

3.7 Proofs in SMT

Pascal Fontaine (LORIA & INRIA – Nancy, FR)

Joint work of Pascal Fontaine, Haniel Barbosa, Jasmin Blanchette, Hans-Jörg Schurr
Main reference Haniel Barbosa, Jasmin Christian Blanchette, Pascal Fontaine: "Scalable Fine-Grained Proofs for Formula Processing", in Proc. of the Automated Deduction – CADE 26 – 26th International Conference on Automated Deduction, Gothenburg, Sweden, August 6-11, 2017, Proceedings, Lecture Notes in Computer Science, Vol. 10395, pp. 398–412, Springer, 2017.
URL http://dx.doi.org/10.1007/978-3-319-63046-5_25

Satisfiability Modulo Theories (SMT) solvers are successfully used for various applications, notably in verification platforms and as back-ends for interactive theorem provers. In both cases, proofs are valuable. In the first, they help to improve confidence, and can also be used for certification in an industrial context. In the second, proofs can be used to reconstruct theorems within the kernel of proof assistants.

Ideally, proofs for SMT should be full and detailed, and at the same time they should not be too large. The overhead of outputting proofs should preferably be small. The talk briefly reviews the various aspects for outputting proofs in SMT, notably in the preprocessing phase, and discusses some issues that still need to be tackled.

3.8 Loop Acceleration for Under-Approximating Program Analysis

Florian Frohn (MPI für Informatik – Saarbrücken, DE)

Joint work of Florian Frohn, Jürgen Giesl, Marc Brockschmidt, Matthias Naaf

In the last years, under-approximating loop acceleration techniques have successfully been used to analyze programs operating on integers. Essentially, such techniques replace single-path loops with non-deterministic straight-line code that under-approximates the effect of the loops. The key to the success of this approach is its ability to construct symbolic under-approximations that cover program traces of arbitrary length. Applications include proving reachability, deducing lower bounds on the worst-case runtime complexity, and proving non-termination.

In this talk, two novel acceleration techniques will be presented. Furthermore, we will discuss several open problems related to loop acceleration. Finally, we will discuss an empirical evaluation of the presented approach.

3.9 From Derivational Complexity to Runtime Complexity of Term Rewriting

Carsten Fuhs (Birkbeck, University of London, GB)

Main reference Carsten Fuhs: “Transforming Derivational Complexity of Term Rewriting to Runtime Complexity”, in Proc. of the Frontiers of Combining Systems – 12th International Symposium, FroCoS 2019, London, UK, September 4-6, 2019, Proceedings, Lecture Notes in Computer Science, Vol. 11715, pp. 348–364, Springer, 2019.
URL http://dx.doi.org/10.1007/978-3-030-29007-8_20

Derivational complexity of term rewriting considers the length of the longest rewrite sequence for arbitrary start terms, whereas runtime complexity restricts start terms to basic terms. Recently, there has been notable progress in automatic inference of upper and lower bounds for runtime complexity. We propose a novel transformation that allows an off-the-shelf tool for inference of upper or lower bounds for runtime complexity to be used to determine upper or lower bounds for derivational complexity as well. Our approach is applicable to derivational complexity problems for innermost rewriting and for full rewriting. We have implemented the transformation in the tool AProVE and conducted an extensive experimental evaluation. Our results indicate that bounds for derivational complexity can now be inferred for rewrite systems that have been out of reach for automated analysis thus far. At the Dagstuhl seminar, we also discussed extensions and possible applications of our approach.

3.10 Decision Procedures Beyond Satisfiability

Jürgen Giesl (RWTH Aachen, DE)

Joint work of Jürgen Giesl, Florian Frohn, Peter Giesl, Marcel Hark
URL https://doi.org/10.1016/j.ipl.2018.06.012, https://doi.org/10.1007/978-3-030-25543-5_24, https://doi.org/10.1007/978-3-030-29436-6_16
Main reference Florian Frohn, Jürgen Giesl: “Constant runtime complexity of term rewriting is semi-decidable”, Inf. Process. Lett., Vol. 139, pp. 18–23, 2018.
URL http://dx.doi.org/10.1016/j.ipl.2018.06.012
Main reference Florian Frohn, Jürgen Giesl: “Termination of Triangular Integer Loops is Decidable”, in Proc. of the Computer Aided Verification – 31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part II, Lecture Notes in Computer Science, Vol. 11562, pp. 426–444, Springer, 2019.
URL http://dx.doi.org/10.1007/978-3-030-25543-5_24
Main reference Jürgen Giesl, Peter Giesl, Marcel Hark: “Computing Expected Runtimes for Constant Probability Programs”, in Proc. of the Automated Deduction – CADE 27 – 27th International Conference on Automated Deduction, Natal, Brazil, August 27-30, 2019, Proceedings, Lecture Notes in Computer Science, Vol. 11716, pp. 269–286, Springer, 2019.
URL http://dx.doi.org/10.1007/978-3-030-29436-6_16

We give an overview on our recent work on finding (sub-)classes of programs with decidable termination or complexity properties. All of our decision procedures give results that go beyond binary “yes/no” answers.

Our first result is that it is semi-decidable whether the runtime complexity of a term rewrite system is constant. In case of constant runtime complexity, our semi-decision procedure also computes the exact worst-case runtime.

In our second result, we consider affine integer programs and show that termination is decidable for programs that consist of a single loop where the update matrix is triangular. Moreover, our procedure can also compute witnesses for eventual non-termination.

Finally, we regard a class of probabilistic programs with constant probabilities (so-called CP programs) and present a procedure to decide (positive) almost sure termination and to compute asymptotically tight bounds on their expected runtime. Based on this, we developed an algorithm to infer the exact expected runtime of any CP program.

3.11 Spacer on Jupyter

Arie Gurfinkel (University of Waterloo, CA)

Joint work of Arie Gurfinkel, Nikolaj Bjørner

Constrained Horn Clauses (CHC) is a fragment of First Order Logic modulo constraints that captures many program verification problems as constraint solving. Safety verification of sequential programs, modular verification of concurrent programs, parametric verification, and modular verification of synchronous transition systems are all naturally captured as a satisfiability problem for CHC modulo theories of arithmetic and arrays.

Of course, satisfiability of CHC is undecidable. Thus, solving them is a mix of science, art, and a dash of magic. In this talk, we have presented a tutorial on using a CHC solver Spacer that is build into SMT solver Z3. The tutorial is illustrated by a Jupyter notebook that shows how different verification problems are modeled by CHC and solved by Spacer.

The corresponding notebook is available at: https://spacerexamples-ariegurfinkel.notebooks.azure.com/

3.12 Abstract Execution

Reiner Hähnle (TU Darmstadt, DE)

Joint work of Reiner Hähnle, Steinhöfel, Dominic
Main reference Dominic Steinhöfel, Reiner Hähnle: "Abstract Execution", in Proc. of the Formal Methods – The Next 30 Years – Third World Congress, FM 2019, Porto, Portugal, October 7-11, 2019, Proceedings, Lecture Notes in Computer Science, Vol. 11800, pp. 319–336, Springer, 2019.
URL http://dx.doi.org/10.1007/978-3-030-30942-8_20

We propose a new static software analysis principle called Abstract Execution, generalizing Symbolic Execution: While the latter analyzes all possible execution paths of a specific program, abstract execution analyzes a partially unspecified program by permitting abstract symbols representing unknown contexts. For each abstract symbol, we faithfully represent each possible concrete execution resulting from its substitution with concrete code. There is a wide range of applications of abstract execution, especially for verifying relational properties of schematic programs. We implemented abstract execution in a deductive verification framework and proved correctness of eight well-known statement-level refactoring rules, including two with loops. For each refactoring we characterize the preconditions that make it semantics-preserving. Most preconditions are not mentioned in the literature.

3.13 Reasoning about Expected Runtimes of Probabilistic Programs (and Quantitative Separation Logic)

Benjamin Kaminski (RWTH Aachen, DE)

Joint work of Benjamin Kaminski, Kevin Batz, Joost-Pieter Katoen, Christoph Matheja, and Thomas Noll
Main reference Benjamin Lucien Kaminski, Joost-Pieter Katoen, Christoph Matheja, Federico Olmedo: "Weakest Precondition Reasoning for Expected Runtimes of Randomized Algorithms", J. ACM, Vol. 65(5), pp. 30:1–30:68, 2018.
URL http://dx.doi.org/10.1145/3208102

We present a weakest-precondition-style calculus a la Dijkstra tailored to reasoning about the expected runtime of probabilistic programs. We put a particular focus on inductive invariant-style reasoning for loops. These invariants are *quantitative* and thus lie inherently beyond the scope of Boolean predicates. Major problems in this context are e.g. to (a) synthesize inductive invariants completely from scratch or (b) shape non-inductive candidates into an inductive invariants.

We also present a quantitative separation logic for reasoning about quantitative aspects of randomized programs that have access to dynamic memory.

3.14 Algorithmic Proof Analysis by CERES

Alexander Leitsch (TU Wien, AT)

Proofs are more than just validations of theorems; they can contain interesting mathematical information like bounds or algorithms. However this information is frequently hidden and proof transformations are required to make it explicit. One such transformation on proofs in sequent calculus is cut-elimination (i.e. the removal of lemmas in formal proofs to obtain proofs made essentially of the syntactic material of the theorem). In his famous paper "Untersuchungen über das logische Schließen" Gentzen showed that for cut-free proofs of prenex end-sequents Herbrand's theorem can be realized via the midsequent theorem. In fact Gentzen defined a transformation which, given a cut-free proof p of a prenex sequent S, constructs a cut-free proof p' of a sequent S' (a so-called Herbrand sequent) which is propositionally valid and is obtained by instantiating the quantifiers in S. These instantiations may contain interesting and compact information on the validity of S. Generally, the construction of Herbrand sequents requires cut-elimination in a given proof p (or at least the elimination of quantified cuts). The method CERES (cut-elimination by resolution) [3] provides an algorithmic tool to compute a Herbrand sequent S' from a proof p (with cuts) of S even without constructing a cut-free version of p [5]. A prominent and crucial principle in mathematical proofs is mathematical induction. However, in proofs with mathematical induction Herbrand's theorem typically fails. To overcome this problem we replace induction by recursive definitions and proofs by proof schemata [1], [2], [4]. An extension of CERES to proof schemata (CERESs) allows to compute inductive definitions of Herbrand expansions (so-called Herbrand systems) representing infinite sequences of Herbrand sequents, resulting in a form of Herbrand's theorem for inductive proofs.

References

1 A. Leitsch, N. Peltier, D. Weller. CERES for first-order schemata. LogCom: Journal of Logic and Computation, vol. 27/7 1897-1954 (2017).

2 C. Dunchev, A. Leitsch, M. Rukhaia, D. WellerCut-Elimination and Proof Schemata. In Logic, Language and Computation, 117-136, Lecture Notes in Computer Science 8984 (2015).

3 M. Baaz and A. Leitsch. Methods of Cut-Elimination, Trends in Logic vol. 34, Springer, (2011) .

4 D. Cerna and A. Leitsch. Schematic Cut Elimination and the Ordered Pigeonhole Principle. IJCAR 2016: 241-256 (2016).

5 A. Leitsch and A. Lolic. Extraction of Epansion Trees. J.Autom.Reasoning 63(1): 95-126 (2019)

3.15 Efficient SAT-Based Reasoning Beyond NP

Joao Marques Silva (Federal University – Toulouse, FR)

The performance improvements made to SAT solvers over the last two decades have reshaped the organization of reasoners for different problem domains, within and beyond NP. This talk provides an overview of recent work on tackling decision and related problems beyond NP. A few successful examples include Maximum Satisfiability (MaxSAT), Propositional Abduction, Quantified Boolean Formulas (QBF), among others.

3.16 Correct-by-Decision Solving and Applications

Alexander Nadel (Intel Israel – Haifa, IL)

Main reference Alexander Nadel: "A Correct-by-Decision Solution for Simultaneous Place and Route", in Proc. of the Computer Aided Verification – 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part II, Lecture Notes in Computer Science, Vol. 10427, pp. 436–452, Springer, 2017.
URL http://dx.doi.org/10.1007/978-3-319-63390-9_23

To reduce a problem, provided in a human language, to constraint solving, one normally maps it to a set of constraints, written in the language of a suitable logic. We highlight a different paradigm, called Correct-by-Decision (CBD), in which the original problem is converted into a set of constraints and a decision strategy, where the decision strategy is essential for guaranteeing the correctness of the modeling. We have successfully applied CBD at Intel for designing scalable SAT-based solutions for several sub-stages of the Physical Design stage of chip design [1, 2, 3]. During our talk, we will walk through an example CBD application for solving the problem of routing under constraints and discuss some open questions, related to CBD.

References

1 Amit Erez and Alexander Nadel. Finding bounded path in graph using SMT for automatic clock routing. In Daniel Kroening and Corina S. Pasareanu, editors, *Computer Aided Verification – 27th International Conference, CAV 2015, San Francisco, CA, USA, July 18-24, 2015, Proceedings, Part II*, volume 9207 of *Lecture Notes in Computer Science*, pages 20–36. Springer, 2015.

2 Alexander Nadel. Routing under constraints. In Ruzica Piskac and Muralidhar Talupur, editors, *2016 Formal Methods in Computer-Aided Design, FMCAD 2016, Mountain View, CA, USA, October 3-6, 2016*, pages 125–132. IEEE, 2016.

3 Alexander Nadel. A correct-by-decision solution for simultaneous place and route. In Rupak Majumdar and Viktor Kuncak, editors, *Computer Aided Verification – 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part II*, volume 10427 of *Lecture Notes in Computer Science*, pages 436–452. Springer, 2017.

3.17 A Resolution-Based Calculus for Preferential Logics

Claudia Nalon (University of Brasilia, BR)

Joint work of Claudia Nalon, Dirk Pattinson
Main reference Cláudia Nalon, Dirk Pattinson: "A Resolution-Based Calculus for Preferential Logics", in Proc. of the Automated Reasoning – 9th International Joint Conference, IJCAR 2018, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 14-17, 2018, Proceedings, Lecture Notes in Computer Science, Vol. 10900, pp. 498–515, Springer, 2018.
URL http://dx.doi.org/10.1007/978-3-319-94205-6_33

Preferential logics are part of a family of conditional logics intended for counterfactual reasoning, allowing to infer and to withdraw conclusions in the presence of new facts. Sequent or tableau calculi for such logics are notoriously hard to construct, and often require additional syntactic structure. Various conditional logics require nested sequents, labelled sequents or special transition formulae, together with non-trivial proofs of either semantic completeness or cut elimination. In this talk, we present a recently developed resolution-based calculus for the preferential logic S and argue that its pure syntactic nature makes it well suited for automation.

3.18 Logically Constrained Rewriting over Bit Vectors

Naoki Nishida (Nagoya University, JP)

Joint work of Naoki Nishida, Yoshiaki Kanazawa
Main reference Yoshiaki Kanazawa, Naoki Nishida, and Masahiko Sakai: "On representation of structures and unions in logically constrained rewriting", IEICE Technical Report SS2018-38, IEICE, Vol. 118, No. 385, pp. 67–72, January 1019 (in Japanese).
URL https://www.ieice.org/ken/paper/20190116P1Jg/eng/

Recently, several methods for verifying imperative programs by means of transformations into Term Rewrite Systems (TRSs, for short) have been investigated. In particular, constrained rewrite systems are popular as sources of such transformations, since logical constraints used for modeling control flows can be separated from terms that represent intermediate states of the execution of target programs. In the existing methods, data types that can be used in target programs are restricted to the integers and their one-dimensional arrays [1], and hence we are not allowed to use other primitive data types, structures, or unions.

In this talk, we briefly introduce Logically Constrained TRSs (LCTRSs, for short) over the bit vectors, which are obtained from C programs with structures and unions. Such LCTRSs can be models of automotive embedded systems, and are useful to verify the corresponding programs. Our framework of rewriting induction, a verification method of equivalence of two functions, works for not only LCTRSs over the integers but also LCTRSs over the bit vectors.

References

1 Carsten Fuhs, Cynthia Kop, and Naoki Nishida. Verifying procedural programs via constrained rewriting induction. *ACM Transactions on Computational Logic*, 18(2):14:1–14:50, June 2017.

3.19 Using SMT solvers to reason about firewalls

Ruzica Piskac (Yale University – New Haven, US)

Joint work of William T. Hallahan, Ennan Zhai, Ruzica Piskac
Main reference William T. Hallahan, Ennan Zhai, Ruzica Piskac: "Automated repair by example for firewalls", in Proc. of the 2017 Formal Methods in Computer Aided Design, FMCAD 2017, Vienna, Austria, October 2-6, 2017, pp. 220–229, IEEE, 2017.
URL http://dx.doi.org/10.23919/FMCAD.2017.8102263

This work present a systematic effort that can automatically repair firewalls, using the programming by example approach. We encode firewall behavior as a set of first-order logic formulas. In our approach, after administrators observe undesired behavior in a firewall, they may provide input/output examples that comply with the intended behavior. Based on the given examples, we automatically synthesize new firewall rules for the existing firewall. This new firewall correctly handles packets specified by the examples, while maintaining the rest of the behavior of the original firewall. Through a conversion of the firewalls to SMT formulas, we offer formal guarantees that the change is correct. Our evaluation results from real-world case studies show that our tool can efficiently find repairs.

3.20 Inductive Inference with Recursion Analysis in Separation Logic

Quang Loc Le (Teesside University – Middlesbrough, GB)

Inductive inference is a vital ingredient of a proof system for reasoning about recursive data structures (e.g., lists and trees) and functions. Especially, supporting an automated inductive theorem prover in a substructural logic, e.g. separation logic, is notoriously hard. In this work, we consider inductive inference for entailment problem in separation logic combined with inductive definitions and arithmetic. We present a novel inductive entailment prover where inductive inference is based on both circular reasoning (in the spirit of Brotherston) and mathematical induction (based on Noetherian principle). The essence of our proposal is a recursion analysis for automatically generating induction rules such that inductive proofs could be obtained locally and efficiently. We have implemented our proposal in a prototype tool and evaluated it over a set of challenging entailment problems taken from a recent competition for solvers in separation logic. The experimental results show that our prover is both effective and efficient. Indeed, it outperformed all existing state-of-the-art entailment solvers.

3.21 Probabilistic Symbolic Execution using Separation Logic

Quoc-Sang Phan (Synopsys Inc. – Mountain View, US)

Probabilistic symbolic execution is a technique to calculate the probability that a program reaches a certain point, which is useful in, for example, reliability analysis. So far this technique has been widely used in programs with numerical inputs, but it enjoys little success when the program makes extensive use of dynamically allocated data structures, such as lists and trees. In this talk, I will present our work-in-progress approach to this problem using separation logic.

References

1 Antonio Filieri, Marcelo F. Frias, Corina S. Pasareanu, Willem Visser. *Model Counting for Complex Data Structures*. SPIN 2015.
2 Long H. Pham, Quang Loc Le, Quoc-Sang Phan, Jun Sun, and Shengchao Qin. *Enhancing Symbolic Execution of Heap-based Programs with Separation Logic for Test Input Generation*. ATVA 2019.
3 Long H. Pham, Quang Loc Le, Quoc-Sang Phan, and Jun Sun. *Concolic Testing Heap-Manipulating Programs*. FM 2019.

3.22 Code commutation

Albert Rubio (Complutense University of Madrid, ES)

Joint work of Elvira Albert, Miguel Gómez-Zamalloa, Miguel Isabel, Albert Rubio
Main reference Elvira Albert, Miguel Gómez-Zamalloa, Miguel Isabel, Albert Rubio: "Constrained Dynamic Partial Order Reduction", in Proc. of the Computer Aided Verification – 30th International Conference, CAV 2018, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 14-17, 2018, Proceedings, Part II, Lecture Notes in Computer Science, Vol. 10982, pp. 392–410, Springer, 2018.
URL http://dx.doi.org/10.1007/978-3-319-96142-2_24

Two pieces of code commute if they reach the same state in any of the two orders of execution. They can commute unconditionally if they commute in any possible initial state or conditionally if it is only for a subset of the initial states. We will present the use of this property in the context of the dynamic partial order technique. The property can be analyzed automatically using SMT solvers, but it becomes challenging when the code involved in the two pieces of code is complex.

3.23 Bit-Vector Interpolation and Quantifier Elimination by Lazy Reduction

Philipp Rümmer (Uppsala University, SE)

Joint work of Backeman, Peter; Zeljic, Aleksandar
Main reference Peter Backeman, Philipp Rümmer, Aleksandar Zeljic: "Bit-Vector Interpolation and Quantifier Elimination by Lazy Reduction", in Proc. of the 2018 Formal Methods in Computer Aided Design, FMCAD 2018, Austin, TX, USA, October 30 – November 2, 2018, pp. 1–10, IEEE, 2018.
URL http://dx.doi.org/10.23919/FMCAD.2018.8603023

The inference of program invariants over machine arithmetic, commonly called bit-vector arithmetic, is an important problem in verification. Techniques that have been successful for unbounded arithmetic, in particular Craig interpolation, have turned out to be difficult to generalise to machine arithmetic: existing bit-vector interpolation approaches are based either on eager translation from bit-vectors to unbounded arithmetic, resulting in complicated constraints that are hard to solve and interpolate, or on bit-blasting to propositional logic, in the process losing all arithmetic structure. We present a new approach to bit-vector interpolation, as well as bit-vector quantifier elimination (QE), that works by lazy translation of bit-vector constraints to unbounded arithmetic. Laziness enables us to fully utilise the information available during proof search (implied by decisions and propagation) in the encoding, and this way produce constraints that can be handled relatively easily by existing interpolation and QE procedures for Presburger arithmetic. The lazy encoding is complemented with a set of native proof rules for bit-vector equations and non-linear (polynomial) constraints, this way minimising the number of cases a solver has to consider.

3.24 Forgetting for Computing Snap-Shots of Ontologies: Progress and Challenges

Renate Schmidt (University of Manchester, GB)

Main reference Jieying Chen, Ghadah Alghamdi, Renate A. Schmidt, Dirk Walther, Yongsheng Gao: "Ontology Extraction for Large Ontologies via Modularity and Forgetting", in Proc. of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019, pp. 45–52, ACM, 2019.
URL http://dx.doi.org/10.1145/3360901.3364424

Forgetting is non-standard reasoning technology to restrict the information in a knowledge base by excluding some of the terms and symbols in the signature. My presentation focussed on the use of forgetting for ontology extraction. Because ontologies can be very large, it is useful to have tools that extract and create snap-shots of an ontology. The creation of ontology extracts is an essential operation for the reuse, creation, evaluation, curation, decomposition, integration and general use of ontologies. For example, it allows ontology modellers to create and work with extracts of an ontology that succinctly summarise the information relating to particular terms in the ontology. These could be used to create new smaller ontologies tailor-made for a particular purpose required by a new application of specific vendors.

After a brief introduction of forgetting the presentation discussed a trial of current forgetting tools in an industry collaboration with SNOMED International and the challenges that this research has thrown up.

For the SNOMED use case the ability to produce extracts for very small signatures compared to the signature of the whole ontology (1% or less) was required. The smaller the extract signature, the more work forgetting tools have in order to compute an extract. The research found that what helps is pre-computing a module and then applying forgetting; in addition signature extension was needed to allow modellers to use existing refsets of concept names. A novel workflow was developed consisting of four stages: (a) signature extension, (b) ontology module extraction, (c) forgetting, and (d) feedback by domain experts, which was evaluated on the SNOMED CT and NCIt ontologies. The investigation used three different modularisation approaches (locality-based, semantic and minimal subsumption modularisation) and three forgetting tools (NUI, LETHE and FAME).

Discussion of the various challenges that remain to be addressed has revealed useful suggestions for improving the ontology extraction process and an alternative faster approach.

3.25 Efficient Validation of FOLID Cyclic Induction Reasoning

Sorin Stratulat (University of Lorraine – Metz, FR)

Checking the soundness of the cyclic induction reasoning for first-order logic with inductive definitions (FOLID) may be costly; the standard checking method is decidable but based on a doubly exponential complement operation for Büchi automata. In this talk, I will present a semi-decidable polynomial method whose most expensive steps recall the comparisons with multiset path orderings. In practice, it has been integrated in the Cyclist prover and successfully checked all the proofs generated with the standard method and included in its distribution.

FOLID cyclic proofs may also be hard to certify. Our method helps to represent the cyclic induction reasoning as being well-founded, where the ordering constraints are automatically built from the analysis of the proofs. Hence, it creates a bridge between the two induction reasoning methods and opens the perspective to use the certification methods adapted for well-founded induction proofs.

3.26 Rule-Based Nonmonotonic Reasoning with Probabilities

Andrzej Szalas (University of Warsaw, PL)

Main reference Andrzej Szałas: "Decision-Making Support Using Nonmonotonic Probabilistic Reasoning". In: Czarnowski I., Howlett R., Jain L. (eds) Intelligent Decision Technologies 2019. Smart Innovation, Systems and Technologies, vol 142. Springer, Singapore
URL https://doi.org/10.1007/978-981-13-8311-3_4

The talk will be focused on a decision-making support by rule-based nonmonotonic reasoning enhanced with probabilities. As a suitable rule-based tool we will analyze Answer Set Programming (ASP) and explore its probabilistic extension permitting the use of probabilistic expressions of two types. The first type represents an externally given prior probability distribution on literals in an answer set program P. The second type represents a posterior distribution conditioned on individual decisions and choices made, together with their consequences represented by answer sets of P.

The ability to compare aspects of both the prior and posterior probabilities in the language of the program P has interesting uses in filtering solutions/decisions one is interested in. A formal characterization of this probabilistic extension to ASP as well as some examples demonstrating its potential use will also be discussed.

The discussed techniques do not increase the complexity of standard ASP-based reasoning.

3.27 A Fixpoint Logic and Dependent Effects for Temporal Property Verification

Tachio Terauchi (Waseda University – Tokyo, JP)

Joint work of Yoji Nanjo, Hiroshi Unno, Eric Koskinen, Tachio Terauchi
Main reference Yoji Nanjo, Hiroshi Unno, Eric Koskinen, Tachio Terauchi: "A Fixpoint Logic and Dependent Effects for Temporal Property Verification", in Proc. of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2018, Oxford, UK, July 09-12, 2018, pp. 759–768, ACM, 2018.
URL http://dx.doi.org/10.1145/3209108.3209204

Existing approaches to temporal verification of higher-order functional programs have either sacrificed compositionality in favor of achieving automation or vice-versa. In this paper we present a dependent-refinement type & effect system to ensure that welltyped programs satisfy given temporal properties, and also give an algorithmic approach—based on deductive reasoning over a fixpoint logic–to typing in this system. The first contribution is a novel type-and-effect system capable of expressing dependent temporal effects, which are fixpoint logic predicates on event sequences and program values, extending beyond the (non-dependent) temporal effects used in recent proposals. Temporal effects facilitate compositional reasoning whereby the temporal behavior of program parts are summarized as effects and combined to form those of the larger parts. As a second contribution, we show that type checking and typability for the type system can be reduced to solving first-order fixpoint logic constraints. Finally, we present a novel deductive system for solving such constraints. The deductive system consists of rules for reasoning via invariants and well-founded relations, and is able to reduce formulas containing both least and greatest fixpoints to predicate-based reasoning.

3.28 Abduction in DL by translation to FOL

Sophie Tourret (MPI für Informatik – Saarbrücken, DE) and Christoph Weidenbach (MPI für Informatik – Saarbrücken, DE)

Joint work of Sophie Tourret, Patrick Koopman, Christoph Weidenbach

Description Logics (DL) are the languages of choice to reason about real world knowledge as represented in web ontologies. One recurrent issue with ontologies is their incompleteness. Abductive reasoning is one way to repair such faulty systems by automatically computing possible explanations for observations that, although not entailed by the ontology, do not contradict it. Most existing abduction techniques for DL have a limited expressiveness and do not scale well. By relying on state-of-the-art tools for abduction in first-order logic, based on SMT, we want to improve on the current situation.

3.29 On Hierarchical Symbol Elimination and Applications

Viorica Sofronie-Stokkermans

Joint work of Viorica Sofronie-Stokkermans, Dennis Peuter

We present possibilities of symbol elimination in extensions of a theory T_0 with additional function symbols whose properties are axiomatised using a set of clauses which we established in [1] and [2]. We analyze situations in which we can perform such tasks in a hierarchical way, relying on existing mechanisms for symbol elimination in T_0. This is for instance possible if the base theory T_0 allows quantifier elimination. We discuss possibilities of extending such methods to situations in which the base theory does not allow quantifier elimination but has a model completion which does (or, in some cases, has a co-theory in which symbol elimination is possible). We discuss the way these results can be used e.g. for abduction, interpolant generation (cf. e.g. [1], [2]), and invariant strengthening [3].

References

1 V. Sofronie-Stokkermans. On Interpolation and Symbol Elimination in Theory Extensions N. Olivetti and A. Tiwari (eds), *Automated Reasoning – 8th International Joint Conference, IJCAR 2016*, LNCS 9706, pages 273–289, Springer (2016)

2 V. Sofronie-Stokkermans. On Interpolation and Symbol Elimination in Theory Extensions. Logical Methods in Computer Science 14(3) (2018)

3 D. Peuter and V. Sofronie-Stokkermans. On Invariant Synthesis for Parametric Systems. In: Fontaine P (ed.), *Automated Deduction – CADE 27 – 27th International Conference on Automated Deduction*, LNCS 11716, pages 385–405, Springer (2019).

3.30 Saturation Theorem Proving: From Inference Rules to Provers

Uwe Waldmann (MPI für Informatik – Saarbrücken, DE)

Joint work of Uwe Waldmann, Jasmin Blanchette, Simon Robillard, Sophie Tourret

One of the indispensable operations of realistic saturation theorem provers is (backward and forward) deletion of subsumed formulas. In presentations of proof calculi, however, subsumption deletion is usually discussed only informally, and in the rare cases where there is a formal exposition, it is typically clumsy. The main reason for this is the fact that the well-known equivalence of dynamic and static refutational completeness holds only for derivations where all deleted formulas are redundant, but using a standard notion of redundancy, a clause C does not make an instance Cσ redundant.

We are working on a generic framework for formal refutational completeness proofs of abstract provers that implement saturation proof calculi. The framework relies on a modular extension of lifted redundancy criteria, which in the end permits not only to cover subsumption deletion, but to model entire prover architectures in such a way that the static refutational completeness of a calculus immediately implies the dynamic refutational completeness of, say, an Otter loop or Discount loop prover implementing the calculus.

A formal proof in Isabelle is currently under development.

3.31 Loop Detection by Logically Constrained Rewriting

Sarah Winkler (University of Verona, IT)

Joint work of Naoki Nishida, Sarah Winkler
Main reference Naoki Nishida, Sarah Winkler: "Loop Detection by Logically Constrained Term Rewriting", in Proc. of the Verified Software. Theories, Tools, and Experiments – 10th International Conference, VSTTE 2018, Oxford, UK, July 18-19, 2018, Revised Selected Papers, Lecture Notes in Computer Science, Vol. 11294, pp. 309–321, Springer, 2018.
URL http://dx.doi.org/10.1007/978-3-030-03592-1_18

Logically constrained rewrite systems (LCTRSs) constitute a very general rewriting formalism that captures simplfication processes in various domains, as well as computation in imperative programs. In both of these contexts, nontermination is a critical source of errors. This talk discusses loop criteria for LCTRSs that are implemented in the tool Ctrl. The usefulness of these criteria is illustrated by applications in four domains: checking (a) LLVM peephole optimizations as well as (b) simplification rules of SMT solvers for potential loops, (c) detecting looping executions of C programs, and (d) establishing nontermination of integer transition systems.

Participants

- Alexander Bentkamp
Free University Amsterdam, NL
- Nikolaj S. Bjørner
Microsoft Research –
GRedmond, US
- Maria Paola Bonacina
Università degli Studi di
Verona, IT
- Florent Capelli
INRIA Lille, FR
- Warren Del-Pinto
University of Manchester, GB
- Rayna Dimitrova
University of Leicester, GB
- Pascal Fontaine
LORIA & INRIA – Nancy, FR
- Florian Frohn
MPI für Informatik –
Saarbrücken, DE
- Carsten Fuhs
Birkbeck, University of
London, GB
- Jürgen Giesl
RWTH Aachen, DE
- Alberto Griggio
Bruno Kessler Foundation –
Trento, IT
- Arie Gurfinkel
University of Waterloo, CA
- Reiner Hähnle
TU Darmstadt, DE
- Matthias Heizmann
Universität Freiburg, DE
- Benjamin Kaminski
RWTH Aachen, DE
- Laura Kovács
TU Wien, AT
- Quang Loc Le
Teesside University –
Middlesbrough, GB
- Alexander Leitsch
TU Wien, AT
- Anthony W. Lin
TU Kaiserslautern, DE
- Joao Marques-Silva
Federal University –
Toulouse, FR
- David Monniaux
VERIMAG – Grenoble, FR
- Alexander Nadel
Intel Israel – Haifa, IL
- Claudia Nalon
University of Brasilia, BR
- Naoki Nishida
Nagoya University, JP
- Quoc Sang Phan
Synopsys Inc. –
Mountain View, US
- Ruzica Piskac
Yale University – New Haven, US
- Albert Rubio
Complutense University of
Madrid, ES
- Philipp Rümmer
Uppsala University, SE
- Andrey Rybalchenko
Microsoft Research –
Cambridge, GB
- Renate Schmidt
University of Manchester, GB
- Martina Seidl
Johannes Kepler Universität
Linz, AT
- Viorica Sofronie-Stokkermans
Universität Koblenz-Landau, DE
- Sorin Stratulat
University of Lorraine –
Metz, FR
- Andrzej Szalas
University of Warsaw, PL
- Tachio Terauchi
Waseda University – Tokyo, JP
- Cesare Tinelli
University of Iowa –
Iowa City, US
- Sophie Tourret
MPI für Informatik –
Saarbrücken, DE
- Andrei Voronkov
University of Manchester, GB &
EasyChair
- Uwe Waldmann
MPI für Informatik –
Saarbrücken, DE
- Christoph Weidenbach
MPI für Informatik –
Saarbrücken, DE
- Thomas Wies
New York University, US
- Sarah Winkler
University of Verona, IT

Report from Dagstuhl Seminar 19381

Application-Oriented Computational Social Choice

Edited by

Umberto Grandi[1], Stefan Napel[2], Rolf Niedermeier[3], and Kristen Brent Venable[4]

1 **University Toulouse Capitole, FR,** umberto.grandi@ut-capitole.fr
2 **Universität Bayreuth, DE,** stefan.napel@uni-bayreuth.de
3 **TU Berlin, DE,** rolf.niedermeier@tu-berlin.de
4 **IHMC – Pensacola, US,** bvenable@ihmc.us

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19381 “Application-Oriented Computational Social Choice”. The seminar was organised around four focus topics: group recommender systems, fair allocation, electoral systems, and interactive democracy. For each topic, an invited survey was given by one of the participants. 26 participants presented their research in a regular talk, and two rump sessions allowed other participants to present their ongoing work and open problems in short talks. A special session was dedicated to software demonstrations, and 3 voting experiments were run during the seminar, also thanks to a mobile experimental laboratory that was brought to Dagstuhl. Finally, three afternoons were dedicated to group works.

Seminar September 15–20, 2019 – http://www.dagstuhl.de/19381
2012 ACM Subject Classification Theory of computation → Algorithmic game theory and mechanism design
Keywords and phrases ai for the social good, collective decision making, multi-agent systems, social choice
Digital Object Identifier 10.4230/DagRep.9.9.45
Edited in cooperation with Arianna Novaro, Dominik Peters

1 Executive Summary

Umberto Grandi (University Toulouse 1 Capitole – FR)
Stefan Napel (Universität Bayreuth – DE)
Rolf Niedermeier (TU Berlin – DE)
K. Brent Venable (IHMC – US)

Computational social choice (COMSOC) combines models from political science and economics with techniques from computer science, to analyze collective decision processes from a computational perspective. Classical contributions include the study of the computational barriers to various forms of manipulation in elections, the definition of novel procedures for distributed resources among a group of human or artificial agents, as well as the study of complex collective decisions such as multi-winner voting rules and voting in combinatorial domains. COMSOC is a thriving field of research, with an international bi-annual workshop now at its 7th edition and a handbook published in 2016 which structures more than a decade of research, but future success will depend on the practical applicability of its findings.

The purpose of this seminar was to address this challenge by stimulating application-driven research in computational social choice, i.e., theoretical studies modeling existing practical problems in all their complexity.

Four areas of COMSOC, which have already proven or bear particular potential for synergies and applicability to real-life problems, were identified as the focus of the seminar. Each of these areas addresses present-day challenges that provide an opportunity for an interdisciplinary approach building on contributions from computer scientists, economists, mathematicians, and political scientists:

- **Recommender systems** is a very successful application that combines several artificial intelligence techniques. Indeed, there have been few other examples of autonomous reasoning tools with comparable impact and pervasiveness in practice.
- **Fair division** has already proven a successful testbed for the application of theoretical work, thanks for the recently launched Spliddit webpage, which provides a user-friendly implementation for a number of algorithms in this field. This experience poses a number of questions and challenges for application-oriented research in fair division and beyond, such as data collection and analysis, possibly leading to new theoretical problems.
- **Interactive democracy** comprises a variety of approaches to make democratic processes more engaging and responsive. For instance, successful design and implementation of online decision platforms presents a multidisciplinary research challenge.
- **Real electoral systems** often have features that are absent in the single or multi-winner systems analyzed in textbooks and scientific papers. Voting theory and computational methods can help to identify non-monotonicity problems of real electoral systems, to provide normative benchmarks for institutional design, and to conduct influence and performance comparisons of different voting arrangements.

The Dagstuhl Seminar 19381 “Application-Oriented Computational Social Choice” brought together 46 invited participants of 15 different nationalities from 4 different continents, with three additional participants choosing to attend our seminar before participating to the Heidelberg Laureate Forum. The list of participants included researchers in Computer Science, Economics, and Political Science, three researchers from the industry (Microsoft, IBM, WinSet Group), and a lab technician.

For each of the focus topics described above, a 1-hour survey was prepared by one of the participants, obtaining an up-to-date overview of current research in the field and its main open problems. Each survey was scheduled on a different day, with 26 regular talks by participants complementing them in the program. Two rump sessions at the beginning of the week allowed a number of the participants to present recent findings, open problems and on-going research in a quick and informal way, stimulating the discussion for the rest of the week.

Given the focus of the seminar on application-oriented research, a special session was dedicated to the presentation of software developed by researchers participating to the seminar. Voting platforms were presented (Whale[1] and OPRA[2]), a library for preference data (Preflib[3]), a platform for online deliberation and consensus building (Vilfredo[4]), as well as a number of tools to support experimental research in social choice. Moreover, the seminar hosted three live voting experiments during the week, two of which used a mobile

[1] https://whale.imag.fr/
[2] https://opra.cs.rpi.edu/polls/main
[3] http://www.preflib.org/
[4] https://www.vilfredo.org/

experimental laboratory that was brought to Dagstuhl thanks to French CNRS and the help of a lab technician from University of Rennes. A detailed report of the experiments and an abstract of all the talks can be found below.

At the beginning of the week short sessions were reserved for individual self-introductions and for the proposition of potential group work. The organisers chose not to organize groups in advance, but to let them form in an iterative fashion during the seminar. A number of proposals were first made, then discussed and adapted, before participants signed up for specific group sessions. A total of 6 hours during the week was dedicated to group works, which led to significant advancements – a detailed report can be read below.

Overall, judging both from anecdotal personal feedback as well as the official results from the anonymous "Survey for Dagstuhl Seminar 19381" (with a median score of 10 out of 11 on the summary question "All in all, how do you rate the scientific quality of the seminar?" and similarly positive answers on the mix of participants, working atmosphere, etc.), the seminar was a very successful experience. It stimulated an already thriving research field to explore more applied research topics and scout for real-world problems. It allowed researchers to get first hand experience on how to run voting experiments, either on an Internet voting platform or in a laboratory, and allowed them to share their research practices. The work conducted in the groups was overall fruitful, already resulting in some paper drafts under preparation. The few suggestions for improvements mostly related to further broadening the mix of participants (more PhD students and junior researchers, more colleagues from nearby fields) and having a slightly less dense program (shorter talks, more time for work in small groups or unplanned activities).

The organisers wish to thank all the Dagstuhl staff for their professional support, the participants of the seminar for their positive attitude and enthusiasm, and the two collectors for putting together the abstracts that compose this report.

2 Table of Contents

Executive Summary

Abstracts of Invited Surveys

Abstracts of Talks

Working Groups and Voting Experiments

3 Abstracts of Invited Surveys

3.1 Developments in Multi-agent Fair Allocation

Haris Aziz (UNSW – Sydney, AU)

I survey some of the recent developments in fair allocation. I also draw some connections with other strands of social choice where fairness can be an important concern.

3.2 From Computational Social Choice to Digital Democracy

Markus Brill (TU Berlin, DE)

Digital Democracy is an umbrella term that encompasses a variety of approaches to make democratic decision making processes more engaging and interactive by utilizing digital instruments such as online citizen participation platforms. In contrast to traditional democratic systems, such online platforms aim to leverage citizen expertise by providing an open collaborative environment with novel interaction possibilities.

The successful design of digital democracy platforms and procedures presents a multidisciplinary research challenge; a particularly pertinent area is the theory of collective decision making (aka social choice theory).

In this talk, I demonstrate how concepts and techniques from social choice theory can be employed to aid the design of online participation platforms and other digital democracy tools. I argue that insights from computational social choice, an active research area at the intersection of computer science, economics, and political science, are particularly relevant for this endeavor.

3.3 A Walk Down K-Street and Beyond: Case Studies in Spatial Vote Modeling

Joseph Godfrey (WinSet Group, LLC – Falls Church, US)

Joint work of Bernard Grofman, Joseph Godfrey

The talk reviewed a set of “case studies” in the application of spatial vote modeling to “real world” settings, ranging from early efforts to sell a software product to K-street lobbyists, through consulting in both governmental and academic communities.

The case studies highlight issues in model specification, the operationalization of social choice constructs, including the use of revealed preferences to locate ideal points, estimation of the status quo and of proposals, the construction of agent-specific indifference curves, and associated computational challenges. Specifically, computation of the Shapley-Owen voting power index, solution concepts such as the strong point, yolk, win set, and uncovered set, as well as dealing with uncertainty, bounded rationality, and the creative/inductive use of data are discussed.

The final case study, and paper, concerns spatial vote models developed using the Siegel-Simon aspirational approach. There are many situations where voters must choose a single alternative and where both the voters and the alternatives can be characterized as points in a one- or two or more-dimensional policy space. In committees and legislatures, often choice among these alternatives will be done via a decision agenda in which alternatives are eliminated until a choice is made, sometimes requiring a final vote against the status quo. A common form for such an agenda is what has been called by Black (1958) standard amendment procedure, a "king of the hill" procedure in which there is an initial alternative who is paired against another alternative, with the winner of that pairwise contest becoming the new winner, and the processes continuing until either the set of feasible alternatives is exhausted or there is a successful motion for cloture. Beginning with a seminal experiment on five person voting games conducted by Plott and Fiorina (1978), there have been a number of experiments on committee voting games with a potentially infinite set of alternatives embedded in a two dimensional policy space. In games where there is a core, i.e., an alternative which, for an odd number of voters, can defeat each and every other alternative in paired comparison, outcomes at or near the core are chosen, but there is also considerable clustering of outcomes even in games without a core. A major concern of the literature has been to develop models to explain the pattern of that clustering in non-core situations. Here, after reviewing the present state of the art, we offer a new family of models based on the Siegel-Simon aspiration approach, in which voters satisfice by choosing "acceptable" alternative, and the set of outcomes that are considered acceptable by each voter changes as the game continues.

3.4 Recommendations and Wagering: Some Surprising Connections to Social Choice

David Pennock (Microsoft – New York, US)

Joint work of David Pennock, Rupert Freeman, C. Lee Giles, Eric Horvitz, Andreas Krause, Dominik Peters, Jens Witkowski, Jenn Wortman Vaughan

Main reference Rupert Freeman, David M. Pennock, Jennifer Wortman Vaughan: "An Equivalence between Wagering and Fair-Division Mechanisms", in Proc. of the The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 – February 1, 2019, pp. 1957–1964, AAAI Press, 2019.

URL https://doi.org/10.1609/aaai.v33i01.33011957

I present two applied setting with direct connections to social choice: recommender systems and wagering mechanisms. I survey recommender systems, emphasizing the history and big picture of the field, including a remembrance of its co-founder, John Riedl, who died of cancer in 2013. I present how some of the usual voting axioms make sense in the context of recommender systems. I discuss some of the practical issues of deployed systems that go beyond the theory. In Part Two, I present truthful wagering mechanisms and show that they are identical to allocation mechanisms in fair division. Wagering mechanisms also correspond to forecaster selection mechanisms and truthful no-regret machine learning algorithms. Finally, wagering connects in interesting ways to participatory budgeting.

4 Abstracts of Talks

4.1 How Hard Is the Manipulative Design of Scoring Systems?

Dorothea Baumeister (Heinrich-Heine-Universität Düsseldorf, DE)

Joint work of Dorothea Baumeister, Tobias Hogrebe
Main reference Dorothea Baumeister, Tobias Hogrebe: "How Hard Is the Manipulative Design of Scoring Systems?", in Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, pp. 74–80, ijcai.org, 2019.
URL http://dx.doi.org/10.24963/ijcai.2019/11

In an election, votes are often given as ordered lists over candidates. A common way of determining the winner is then to apply some scoring system, where each position is associated with a specific score. This setting is also transferable to other situations, such as sports tournaments. The design of such systems, i.e., the choice of the score values, may have a crucial influence on the outcome. We study the computational complexity of two related decision problems. In addition, we provide a case study of data from Formula 1 using ILP formulations. Our results show that under some mild conditions there are cases where the actual scoring system has no influence, whereas in other cases very small changes may lead to a different winner. This may be seen as a measure of robustness of the winning candidate.

4.2 Building an Experiment on Multiwinner Elections

Sylvain Bouveret (University of Grenoble, FR), Jérôme Lang (University Paris-Dauphine, FR), and Vincent Merlin (Caen University, FR)

Joint work of Sylvain Bouveret, Ayça GiritliGil, Jérôme Lang, Vincent Merlin

Since the pioneer works of Plott (1967), and Florina, Morris and Plot (1978), experimental economics, either through lab experiments and or field experiments, contributed to the advancement of research on voting roles, to precise how voters react, adapt their preferences, and vote when confronted to different voting mechanisms. A stream of research aims to understand how voters act strategically in voting and election (Myerson et al., 1993; Laslier et al., 2010). Another direction is to understand how people would react to a modification of the electoral rule (Baujard et al., 2014). A third option is to elicit, under a veil of ignorance, the principles that the voters would back when confronted to a choice. In this line, the major contributions are due to Sertel and Giritligil (2003) and Giritligil and Sertel (2005). These panel studies aim to extract preferences of subjects on how to aggregate individual preferences in a social choice context. Sertel and Giritligil (2003) attempt to empirically understand public preferences concerning four social choice rules of focus, namely Plurality, Plurality with Runoff, the Majoritarian Compromise and the Borda Rule. Giritligil and Sertel (2005), on the other hand, aim to test whether the support for the Borda winner or the Condorcet winner increases when they are among the "Majoritarian Approved" candidates. Recently, researchers working on the axiomatic analysis of committee election rules have emphasized the fact that some voting rules are more suitable in certain context than in others (Faliszewski et al., 2017). They distinguish between three types of contexts. We may wish to select a committee 1) to elect an assembly that represents the preferences of the voters 2) to shortlist a number of candidates, based on their excellence, 3) to get a menu of objects as diverse as possible, so that the tastes of each participant are somehow satisfied. But one may wonder whether these distinctions are pertinent. Our objective in the experiment is to understand the principles that govern the preferences of voters in committee elections in specified and neutral contexts. To do so, we borrow the experimental protocols used by Sertel and Giritligil (2003) and Giritligil and Sertel (2005). In this presentation, we will make the participant of the seminar test the experiment, we will discuss the protocol we are currently working on, and we will comment preliminary results from a pilote.

References

1 Baujard, A., Igersheim, H., Lebon, I., Gavrel, F., Laslier, J.F. 2014. "Who's favored by evaluative voting? An experiment conducted during the 2012 French presidential election", Electoral Studies, 34, 131-145.

2 Faliszewski, P., Skowron, P., Slinko, A., Talmon, N. 2017. "Multiwinner voting: A New Challenge for Social Choice Theory". Trends in Computational Social Choice. Editor: Endriss, U. AI Access.

3 Fiorina, M., Plott, C. R. 1978. "Committee decisions under majority rule: An experimental study", American Political Science Review, 72, 575-598.

4 Forsythe, R., Myerson, R.B., Rietz, T.A., Weber, R.J. 1993. "An experiment on coordination in multi-candidate elections: The importance of polls and election histories", Social Choice and Welfare, 10(3), 223-247.

5 Gehrlein, W.V. 1985. "The Condorcet criterion and committee selection", Mathematical Social Sciences 10(3), 199-209.

6 Giritligil Kara, A. E., Sertel, M. R. 2005. "Does majoritarian approval matter in selecting a social choice rule? An exploratory panel study", Social Choice and Welfare, 25(1), 43-73.

7 Plott, C. R. 1967. "A notion of equilibrium under majority rule", American Economic Review, 57, 787-806. Press.

8 Van der Straeten, K., Laslier, J. F., Sauger, N. 2010. "Strategic, sincere, and heuristic voting under four election rules: An experimental study", Social Choice and Welfare, 35(3), 435-472.

4.3 Simple Characterizations of Approval Voting

Florian Brandl (Stanford University, US) and Dominik Peters (University of Oxford, GB)

Approval voting allows every voter to cast a ballot of approved alternatives and chooses the alternatives with the largest number of approvals. Due to its simplicity and superior theoretical properties it is a serious contender for use in real-world elections. We support this claim by giving seven characterizations of approval voting based on normatively appealing axioms. All our results involve the reinforcement axiom, which requires choices to be consistent across different electorates. In addition, we consider strategyproofness, consistency with majority opinions, consistency under cloning alternatives, and invariance under removing inferior alternatives. We prove our results by reducing them to a single base theorem, for which we give a simple and intuitive proof.

4.4 Efficient and Envy-Free Resource Allocation with Few Agents: Theory and Experiments

Robert Bredereck (TU Berlin, DE) and Rolf Niedermeier (TU Berlin, DE)

Joint work of Robert Bredereck, Andrzej Kaczmarczyk, Dusan Knop, Rolf Niedermeier
Main reference Robert Bredereck, Andrzej Kaczmarczyk, Dusan Knop, Rolf Niedermeier: "High-Multiplicity Fair Allocation: Lenstra Empowered by N-fold Integer Programming", in Proc. of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019, pp. 505–523, ACM, 2019.
URL https://doi.org/10.1145/3328526.3329649

On the theoretical side we study the (parameterized) computational complexity of problems in the context of fair allocations of indivisible goods. More specifically, we show fixed-parameter tractability results for a broad set of problems concerned with envy-free, Pareto-efficient allocations of items (with agent-specific utility functions) to agents. In principle, this implies efficient exact algorithms for these in general computationally intractable problems whenever we face instances with few agents and low maximum (absolute) utility values.

On the practical side, we show that our theoretical framework can be implemented in a practically useful way. As opposed to what theoretical performance guarantees suggest, we are able to solve reasonably large instances within seconds (e.g. all spliddit.org instances). We discuss preliminary experimental results and discuss possible extensions of our framework.

4.5 Hedonic Diversity Games

Edith Elkind (University of Oxford, GB)

Joint work of Edith Elkind, Niclas Boehmer, Robert Bredereck, Ayumi Igarashi
Main reference Robert Bredereck, Edith Elkind, Ayumi Igarashi: "Hedonic Diversity Games", in Proc. of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019, pp. 565–573, 2019.
URL https://dl.acm.org/doi/10.5555/3306127.3331741
Main reference Niclas Boehmer, Edith Elkind: "Individual-Based Stability in Hedonic Diversity Games", CoRR, Vol. abs/1911.08669, 2019.
URL https://arxiv.org/abs/1911.08669

We consider a coalition formation setting where players belong to two types and have preferences over the ratio of the two types in their coalition. We formalize this setting as a hedonic game and consider existence and complexity of finding stable outcomes for various hedonic games solution concepts, such as Nash stability, individual stability and core stability. We consider two ways of extending our model to more than two types of players and provide initial results for this more general setting.

4.6 How Similar Are Two Elections?

Piotr Faliszewski (AGH University of Science & Technology – Krakow, PL)

Joint work of Piotr Faliszewski, Piotr Skowron, Arkadii Slinko, Stanislaw Szufa, Nimrod Talmon
Main reference Piotr Faliszewski, Piotr Skowron, Arkadii Slinko, Stanislaw Szufa, Nimrod Talmon: "How Similar Are Two Elections?", in Proc. of the The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pp. 1909–1916, AAAI Press, 2019.
URL https://doi.org/10.1609/aaai.v33i01.33011909

In this we discuss the problem of measuring distances between elections. We are given two elections, $E_1 = (C_1, V_1)$ and $E_2 = (C_2, V_2)$, where $|C_1| = |C_2|$ and $|V_1| = |V_2|$. That is, both elections have the same number of candidates and the same number of voters, but these are possibly different candidates and voters. In both elections, each voter ranks the candidates from best to worst. We seek a distance that is neutral and anonymous, i.e., which is invariant with respect to permuting the names of the candidates and voters. We introduce a family of what we call *isomorphic distances* which satisfy this condition. Unfortunately, these distances turn out to be computationally very challenging. With one (not so useful) exception, they are NP-hard to compute and hard to approximate with better than linear approximation ratios. While there are FPT algorithms for them, they are quite slow. Then we briefly discuss a very different kind of distance and we show a visual presentation of the distances between elections generated according to a number of classic statistical cultures (including the impartial culture model, Mallows models, urn models, and Euclidean elections).

4.7 Foundations for Liquid Democracy: An Overview

Davide Grossi (University of Groningen, NL)

Liquid democracy is a proxy voting method where proxies are delegable: individual A may delegate her vote to B, who in turn may delegate it to C and so on. Ultimately, an individual's vote carries the weight given by the number of other individuals who (directly or indirectly) entrusted her as proxy. The method has been popularized by the Piratenpartei in Germany as well as other small parties and grassroots organizations worldwide, and it has recently been object of a series of publications in the computational social choice and multi-agent systems communities. In this talk I will overview this stream of work and point to open research directions.

4.8 Fair and Efficient Allocation: Moving Beyond Additivity

Ayumi Igarashi (University of Tokyo, JP)

Joint work of Ayumi Igarashi, Nawal Benabbou, Mithun Chakraborty, Yair Zick

We present new results on the fair and efficient allocation of indivisible goods to agents that have monotone, submodular, non-additive valuation functions over bundles. In particular, we show that, if such a valuation function additionally has binary marginal gains, a socially optimal (i.e. utilitarian social welfare-maximizing) allocation that achieves envy-freeness up to one item exists and can be computed efficiently. We also prove that the Nash welfare-maximizing and the leximin allocations both exhibit this fairness-efficiency combination, by showing that they can be achieved by minimizing any symmetric strictly convex function over utilitarian optimal outcomes. Moreover, for a subclass of these valuation functions based on maximum (unweighted) bipartite matching, we show that a leximin allocation can be computed in polynomial time.

4.9 Empirical Evidence for Consensus Among Voting Rules

Christian Klamler (Universität Graz, AT)

Joint work of Christian Klamler, Andreas Darmann, Julia Grundner

We use preference data from the 2015 Styrian parliament election to analyze different voting rules. An exit poll right after the election collected data on ordinal and cardinal preferences from approximately 1000 actual voters. First, we determine the hypothetical social outcomes under different voting rules. Second, we provide a categorization of different types of parties and analyze the impact of certain voting rules on the performances of the parties. In addition, distance measures have been considered to determine the closeness to changes for the different voting rules. It turns out, that despite the similarity of the voting results, certain outcomes are closer to being changed than others. Finally, based on the ranking data, a bootstrap analysis has been performed. In general, most of the conclusions from behavioral social choice about the consensus among voting rules in real-world elections can be confirmed.

4.10 Decisions with a Continuum of Options

Sascha Kurz (Universität Bayreuth, DE)

Joint work of Sascha Kurz, Issofa Moyouwou, Hilaire Touyem
Main reference Sascha Kurz: "Measuring voting power in convex policy spaces", Economies 2(1): 45–77, 2014.
URL https://doi.org/10.3390/economies2010045

The Shapley-Shubik index was designed to evaluate the power distribution in committee systems drawing binary decisions and is one of the most established power indices. It was generalized to decisions with more than two levels of approval in the input and output. In the limit we have a continuum of options. You may think of e.g. tax rates. For these games with interval decisions we prove an axiomatization of a power measure and show that the Shapley-Shubik index for simple games, as well as for (j,k) simple games, occurs as a special discretization. This relation and the closeness of the stated axiomatization to the classical case suggests to speak of the Shapley-Shubik index for games with interval decisions, that can also be generalized to a value. Also for the Penrose-Banzhaf index there exists a variant for games with interval decisions in the literature on aggregation functions. The general framework of games with a continuum of options deserves to be explored more. We collect a list of some open problems in that direction.

4.11 Fairness in Long-Term Group Decision Making

Martin Lackner (TU Wien – Austria)

Joint work of Martin Lackner, Paul Harrenstein, Marie-Louise Lackner
Main reference Paul Harrenstein, Marie-Louise Lackner, Martin Lackner: "A Mathematical Analysis of an Election System Proposed by Gottlob Frege", CoRR, Vol. abs/1907.03643, 2019.
URL http://arxiv.org/abs/1907.03643

I am discussing fairness in long-term group decision making, primarily based on the paper "A Mathematical Analysis of an Election System Proposed by Gottlob Frege" by Paul Harrenstein, Marie-Louise Lackner, and Martin Lackner.

4.12 2018 Participatory Budgeting in Portugalete: a Case Study

Annick Laruelle (University of the Basque Country – Bilbao, ES)

TIn Europe participatory budgeting by cities usually consists of a five-steps procedure: (1) a city announces an amount to be allocated and a voting rule; (2) citizens make proposals for improving the city; (3) the city administration selects a set of feasible projects and assigns a cost to each of them; (4) citizens vote for projects according to the voting rule; (5) the city announces the winning proposals. Those proposals are implemented. The voting step is usually described as a classical multi-winner election. The only economic constraint is that the total cost of the winning projects must not exceed the amount decided by the city. In this talk we a case study is analysed: the 2018 participatory budgeting of the city of Portugalete (Spain). This analysis give insight on the properties that participatory budgeting mechanisms should display.his is a dummy text.

4.13 Life of the Party

Omer Lev (Ben Gurion University – Beer Sheva, IL)

Joint work of Omer Lev, Yoram Bacharach, Allan Borodin, Yoda Lewenberg, Jeffery S. Rosenschein, Nisarg Shah, Tyrone Strangway, Yair Zick
Main reference Allan Borodin, Omer Lev, Nisarg Shah, Tyrone Strangway: "Primarily about Primaries", in Proc. of the The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pp. 1804–1811, AAAI Press, 2019.
URL https://doi.org/10.1609/aaai.v33i01.33011804
Main reference Allan Borodin, Omer Lev, Nisarg Shah, Tyrone Strangway: "Big City vs. the Great Outdoors: Voter Distribution and How It Affects Gerrymandering", in Proc. of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, pp. 98–104, ijcai.org, 2018.
URL https://doi.org/10.24963/ijcai.2018/14
Main reference Yoram Bachrach, Omer Lev, Yoad Lewenberg, Yair Zick: "Misrepresentation in District Voting", in Proc. of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, pp. 81–87, IJCAI/AAAI Press, 2016.
URL https://www.ijcai.org/Abstract/16/019

When a decision making process is composed of sub-units which have their own internal decision-making processes, such as political parties, it is useful to consider this as a multi-staged electoral process. In particular, we discuss two settings:

1. Primaries, in which 2 political parties (composed of a sizable part of the electorate) decide on their candidate, and the two final candidates go on to a the general electorate. We show this process can increase the distortion by O(1) compared with the direct case (in which all candidates run in the general election), while it can decrease the distortion by O(n).
2. District voting, in which each district makes its choice from among all the candidates, and the candidate with the plurality of districts is the winner. We show the price of districting (again, compared to the direct case), and show that geographical distribution matters significantly, even beyond gerrymandering, showing the urban/rural party divide is meaningful in this setting.

4.14 A Computational Framework for Identity

Janelle C. Mason (North Carolina A&T State University – Greensboro, US)

This presentation presents a computational framework for identity (initially concerning a culprit in a crime scene), which draws its background from various sources across multiple disciplines. Situation theory (as articulated by Barwise and Perry and partially formalized by Devlin) provides a background for the structure of information as situations support information and can carry information about other situations. Semantic Web resources are used to capture information and its structure. The resources used include the Web Ontology Language (OWL) for constructing ontologies, the Resource Description Framework (RDF) for encoding scenarios in triple stores per these ontologies, SPARQL Protocol and RDF Query Language (SPARQL) for querying the triple stores, Semantic Web Rule Language (SWRL) for representing rules, and Jena as an overall Semantic Web framework. The aim is to use the structured information in the RDF triple stores as evidence for identity hypotheses regarding human agents. Collaboration is done with the Criminal Justice Department at North Carolina A&T State University as criminal justice specializes in identifying agents. Identity is an equivalence relation and that it is addressed computationally in term rewriting. Given the structured information, Dempster-Shafer theory is used to reason about how evidence for various identity hypotheses combines and is discounted. In producing a case for the identity of an agent, measures of belief indicate how the evidence conspires to support given identity hypotheses, but they do not flesh out a case. For this we use argumentation schemes, which are forms that allow one to identify and evaluate types of argumentation that occur in everyday discourse. Most of the arguments of interest are defeasible, that is, the conclusion can be tentatively accepted given the known evidence but may have to be altered as new evidence is introduced.

4.15 Flexible Representative Democracy: An Introduction with Binary Issues

Nicholas Mattei (Tulane University – New Orleans, US)

Joint work of Nicholas Mattei, Ben Abramowitz
Main reference Ben Abramowitz, Nicholas Mattei: "Flexible Representative Democracy: An Introduction with Binary Issues", in Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, pp. 3–10, ijcai.org, 2019.
URL https://doi.org/10.24963/ijcai.2019/1

We introduce Flexible Representative Democracy (FRD), a novel hybrid of Representative Democracy (RD) and direct democracy (DD), in which voters can alter the issue-dependent weights of a set of elected representatives. In line with the literature on Interactive Democracy, our model allows the voters to actively determine the degree to which the system is direct versus representative. However, unlike Liquid Democracy, FRD uses strictly non-transitive delegations, making delegation cycles impossible, and maintains a fixed set of accountable elected representatives. We present FRD and analyze it using a computational approach with issues that are binary and symmetric; we compare the outcomes of various democratic systems using Direct Democracy with majority voting as an ideal baseline. First, we demonstrate the shortcomings of Representative Democracy in our model. We provide NP-Hardness results for electing an ideal set of representatives, discuss pathologies, and demonstrate empirically that common multi-winner election rules for selecting representatives do not perform well in expectation. To analyze the behavior of FRD, we begin by providing theoretical results on how issue-specific delegations determine outcomes. Finally, we provide empirical results comparing the outcomes of RD with fixed sets of proxies across issues versus FRD with issue-specific delegations. Our results show that variants of Proxy Voting yield no discernible benefit over RD and reveal the potential for FRD to improve outcomes as voter participation increases, further motivating the use of issue-specific delegations.

4.16 Aggregation of Incomplete CP-nets

Arianna Novaro (Paul Sabatier University – Toulouse, FR)

Joint work of Arianna Novaro, Adrian Haret, Umberto Grandi
Main reference Adrian Haret, Arianna Novaro, Umberto Grandi: "Preference Aggregation with Incomplete CP-Nets", in Proc. of the Principles of Knowledge Representation and Reasoning: Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October - 2 November 2018, pp. 308–318, AAAI Press, 2018.
URL https://aaai.org/ocs/index.php/KR/KR18/paper/view/18069

In this short talk I presented the framework of incomplete, or generalized, CP-nets and their aggregation according to different semantics known in the literature (ie., Pareto, max, maj, and rank). A motivating example for such a setting is that of an online booking service where multiple people (eg., a group of friends) may want to provide their conditional preferences in a flexible way. From a computational complexity perspective, most of the dominance and optimality problems in this setting turn out to be as hard as their single-agent counterpart (PSPACE).

4.17 Pie-Chart Voting: An Annotated Bibliography

Dominik Peters (University of Oxford, GB)

A group of agents owns a common resource (their budget) and needs to decide how to split the budget among various projects. If the projects can receive an arbitrary level of funding (and the budget is perfectly divisible), then the result of the decision process can be visualized as a pie chart. I give a survey of recent work on this topic, focussing on fairness and on strategic issues. The presentation is guided by different possible applications, and which aggregation rules are suitable for them.

4.18 Learning rankings for job recommendations (with constraints)

Karishma Rajesh Sharma (USC – Los Angeles, US)

Joint work of Karishma Rajesh Sharma, Kuan Liu, Anoop Kumar, Prem Natarajan
URL https://github.com/ksharmar/Recsys17

This is part of a rump session.

We address the problem of learning rankings in recommender systems under cold start and constrained settings. We presented our solution to the job recommendation task for ACM RecSys 2017 challenge. The two key challenges addressed are (1) cold start – we need to recommend new items (jobs) with no history of preferences from users (2) constraints – we want to limit the number of recommendations pushed to a user and also ensure maximum utility (relevance) of recommendations to both users and items (companies posting the jobs). We propose a matrix factorization approach for collaborating filtering and incorporate content embeddings to address the cold start items. The model is formulated to obtain preferences (rankings) of users on items and preferences (rankings) of items on users. We then utilize the stable matching algorithm to select optimal recommendation pairs that maximize preferences of both users and items and limit the number of recommendations pushed to a user. The details of the working paper and implementation are available at https://github.com/ksharmar/Recsys17.

4.19 A Framework for Approval-Based Budgeting Methods

Nimrod Talmon (Ben Gurion University – Beer Sheva, IL) and Piotr Faliszewski (AGH University of Science & Technology – Krakow, PL)

Main reference Nimrod Talmon, Piotr Faliszewski: "A Framework for Approval-Based Budgeting Methods", in Proc. of the The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pp. 2181–2188, 2019.
URL http://dx.doi.org/10.1609/aaai.v33i01.33012181

We define and study a general framework for approval-based budgeting methods and compare certain methods within this framework by their axiomatic and computational properties. Furthermore, we visualize their behavior on certain Euclidean distributions and analyze them experimentally.

4.20 Capacitated Facility Location Problems

Toby Walsh (UNSW – Sydney, AU)

I consider the facility location problem in the one-dimensional setting where each facility can serve a limited number of agents from the algorithmic and mechanism design perspectives. From the algorithmic perspective, the optimization problem, where the goal is to locate facilities to minimize either the total cost to all agents or the maximum cost of any agent is NP-hard. I also consider the problem from a mechanism design perspective where the agents are strategic and need not reveal their true locations. We show that several natural mechanisms for the uncapacitated setting either lose strategyproofness or a bound on the solution quality for the total or maximum cost objective when applied to the capacitated problem. I discuss a new mechanisms that is strategyproof and achieves approximation guarantees that almost match the lower bounds.

4.21 A Mathematical Model For Optimal Decisions In A Representative Democracy

Lirong Xia (Rensselaer Polytechnic Institute – Troy, US)

Joint work of Lirong Xia, Malik Magdon-Ismail
Main reference Malik Magdon-Ismail, Lirong Xia: "A Mathematical Model For Optimal Decisions In A Representative Democracy", in Proc. of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada, pp. 4707–4716, 2018.
URL http://papers.nips.cc/paper/7720-a-mathematical-model-for-optimal-decisions-in-a-representative-democracy

Direct democracy, where each voter casts one vote, fails when the average voter competence falls below 50%. This happens in noisy settings when voters have lim- ited information. Representative democracy, where voters choose representatives to vote, can be an elixir in both these situations. We introduce a mathematical model for studying representative democracy, in particular understanding the parameters of a representative democracy that gives maximum decision making capability. Our main result states that under general and natural conditions, (1) for fixed voting cost, the optimal number of representatives is linear; (2) for polynomial cost, the optimal number of representatives is logarithmic.

5 Working Groups and Voting Experiments

5.1 Democratic Decision-Making, Deliberation and Blockchain

Davide Grossi (University of Groningen, NL) and Ehud Shapiro (Weizmann Institute – Rehovot, IL)

The group was structured around two main topics: (1) the interface between blockchain protocols and computational social choice; (2) models of deliberative processes that could provide insights for the design of deliberation support platforms (such as Vilfredo, www.vilfredo.org). The discussion of topic (1) managed to provide an inventory of open issues in current blockchain research which may benefit from insights from computational social choice and more broadly economic theory, such as: committee selection during consensus; genuine individual identifiers to dispense with proof-of-work as a method for Sybil resistance; governance (or 'voting on how to vote'); the design of suitable token systems (cryptocurrencies). The discussion of topic (2) managed to focus on a preliminary model representing deliberation as a non-deterministic process of coalition formation on a metric space. We are confident this approach can lead to interesting results and novel insights.

5.2 Outreach Activities and Real-Life Experiments

Annick Laruelle (University of the Basque Country – Bilbao, ES) and Arianna Novaro (Paul Sabatier University – Toulouse, FR)

The group discussed three main topics: (i) outreach activities for the dissemination of COMSOC results, (ii) real-life experiments in participatory budgeting and voting, and (iii) resources and ideas for teaching COMSOC. Members of the group shared their personal experiences and ideas on the three topics, leading to the suggestion of implementing an online platform where members of the COMSOC community could share their resources on teaching, outreach and real-life experiments. Pros and cons of possible implementation designs were also discussed. During the discussion it was also mentioned the possibility of interacting with other fields (e.g., systems engineering) where COMSOC techniques could be directly applied. Finally, innovative teaching practices for explaining voting rules to an audience of non-experts were outlined at the end of the final group session.

5.3 Social Choice With Uncertain Preferences

Lirong Xia (Rensselaer Polytechnic Institute – Troy, US), Haris Aziz (UNSW – Sydney, AU), Edith Elkind (University of Oxford, GB), Piotr Faliszewski (AGH University of Science & Technology – Krakow, PL), Ayumi Igarashi (University of Tokyo, JP), Jérôme Lang (University Paris-Dauphine, FR), Omer Lev (Ben Gurion University – Beer Sheva, IL), Nicholas Mattei (Tulane University – New Orleans, US), Dominik Peters (University of Oxford, GB), Piotr Skowron (University of Warsaw, PL), and Kristen Brent Venable (IHMC – Pensacola, US)

Classical social choice assumes that voters' preferences are represented by a linear order. However, in many situations agents preferences are uncertain. The problem becomes more prominent in modern applications of social choice such as business decision-making, high frequency e-democracy, consensus on blockchains, and many more.

The working group tried to formalize the direction of social choice with uncertain preferences by identifying the following directions (instead of solving open questions):

1. The source of uncertainty can come from: 1. agents are uncertain about their own preferences, 2. the system is uncertain about agents' preferences, or 3. agents are uncertain about other agents' preferences or behavior.
2. Uncertain preferences can be represented by: probabilistic distributions, fuzzy preferences, possibility theory, decision field theory, conjoint measurement
3. The social choice problem can be: (single and multi-winner) voting, participatory budgeting, matching, kidney exchange resource allocation, etc.

The group also discussed issues in preference data collection, group decision support systems, and open-source packages.

5.4 Voting Experiment: Heuristic Strategies in Uncertain Approval Voting Environments

Kris Brent Venable (IHMC – Pensacola, US) and Nicholas Mattei (Tulane University – New Orleans, US)

Joint work of Jaelle Scheuerman, Jason L. Harman, Kristen Brent Venable, Nicholas Mattei
Main reference Jaelle Scheuerman, Jason L. Harman, Nicholas Mattei, Kristen Brent Venable: "Heuristic Strategies in Uncertain Approval Voting Environments", CoRR, Vol. abs/1912.00011, 2019.
URL https://arxiv.org/abs/1912.00011

In many collective decision making situations, agents vote to choose an alternative that best represents the preferences of the group. Agents may manipulate the vote to achieve a better outcome by voting in a way that does not reflect their true preferences. In real world voting scenarios, people often do not have complete information about other voter preferences and it can be computationally complex to identify a strategy that will maximize their expected utility. In such situations, it is often assumed that voters will vote truthfully rather than expending the effort to strategize. However, being truthful is just one possible heuristic that may be used. In this paper, we examine the effectiveness of heuristics in single winner and multi-winner approval voting scenarios with missing votes. In particular, we look at heuristics where a voter ignores information about other voting profiles and makes their decisions based solely on how much they like each candidate. In a behavioral experiment, we show that people vote truthfully in some situations and prioritize high utility candidates in others. We examine when these behaviors maximize expected utility and show how the structure of the voting environment affects both how well each heuristic performs and how humans employ these heuristics.

5.5 Voting Experiment: Multiple Elections in Iterative Voting

Jérôme Lang (University Paris-Dauphine, FR) and Umberto Grandi (University Toulouse Capitole, FR)

Joint work of Jérôme Lang, Ali Ozkes, Umberto Grandi, Stéphane Airiau

This voting experiment was organised on the Thursday of the workshop for 21 participants. A temporary lab was set up in one of the rooms, using a set of tablets connected on a local network set up by lab technician Elven Priour (CNRS and University of Rennes, France).

Participants were confronted with a multiple election: given a 2x2 matrix, decide if the row will be Top or Bottom, and the column Left or Right. Both decisions would be taken separately by majority. However, the utility of each participant-voter depended on the result of both elections combined. It has been shown in theory that similar elections can generate paradoxical situations in which the collective result is not among the most preferred alternatives (in some cases it is the least preferred) for each voter. In this experiment participants were confronted with an iterated setting, in which each vote would be repeated, and the outcomes of previous votes shown to voters, until the winning combination did not change for three turns. What we could monitor in the data is the evolution of the average Borda score of the winning combination (a parameter also known as ASI), as well as the behaviour of voters facing an uncertain and complex election.

Participants

- Haris Aziz
UNSW – Sydney, AU
- Dorothea Baumeister
Heinrich-Heine-Universität Düsseldorf, DE
- Abdelhak Bentaleb
National University of Singapore, SG
- Sylvain Bouveret
University of Grenoble, FR
- Florian Brandl
Stanford University, US
- Felix Brandt
TU München, DE
- Robert Bredereck
TU Berlin, DE
- Markus Brill
TU Berlin, DE
- Jiehua Chen
TU Wien, AT
- Cristina Cornelio
IBM T.J. Watson Research Center – Yorktown Heights, US
- Ronald de Haan
University of Amsterdam, NL
- Edith Elkind
University of Oxford, GB
- Ulle Endriss
University of Amsterdam, NL
- Piotr Faliszewski
AGH University of Science & Technology – Krakow, PL
- Joseph Godfrey
WinSet Group, LLC – Falls Church, US
- Umberto Grandi
University Toulouse Capitole, FR
- Davide Grossi
University of Groningen, NL
- Ayumi Igarashi
University of Tokyo, JP
- Christian Klamler
Universität Graz, AT
- Sascha Kurz
Universität Bayreuth, DE
- Martin Lackner
TU Wien, AT
- Jérôme Lang
University Paris-Dauphine, FR
- Annick Laruelle
University of the Basque Country – Bilbao, ES
- Omer Lev
Ben Gurion University – Beer Sheva, IL
- Andrea Loreggia
University of Padova, IT
- Nicola Frederike Maaser
Aarhus University, DK
- Janelle C. Mason
North Carolina A&T State University – Greensboro, US
- Nicholas Mattei
Tulane University – New Orleans, US
- Nicolas Maudet
Sorbonne University – Paris, FR
- Reshef Meir
Technion – Haifa, IL
- Vincent Merlin
Caen University, FR
- Stefan Napel
Universität Bayreuth, DE
- Rolf Niedermeier
TU Berlin, DE
- Arianna Novaro
Paul Sabatier University – Toulouse, FR
- David Pennock
Microsoft – New York, US
- Dominik Peters
University of Oxford, GB
- Elven Priour
University of Rennes, FR
- Jörg Rothe
Heinrich-Heine-Universität Düsseldorf, DE
- M. Remzi Sanver
University Paris-Dauphine, FR
- Ehud Shapiro
Weizmann Institute – Rehovot, IL
- Karishma Rajesh Sharma
USC – Los Angeles, US
- Piotr Skowron
University of Warsaw, PL
- Arkadii Slinko
University of Auckland, NZ
- Pietro Speroni di Fenizio
Dublin, IE
- Nimrod Talmon
Ben Gurion University – Beer Sheva, IL
- Paolo Turrini
University of Warwick – Coventry, GB
- Kristen Brent Venable
IHMC – Pensacola, US
- Toby Walsh
UNSW – Sydney, AU
- Lirong Xia
Rensselaer Polytechnic Institute – Troy, US

Report from Dagstuhl Seminar 19391

Data Ecosystems: Sovereign Data Exchange among Organizations

Edited by

Cinzia Cappiello[1], Avigdor Gal[2], Matthias Jarke[3], and Jakob Rehof[4]

1 **Polytechnic University of Milan, IT,** `cinzia.cappiello@polimi.it`
2 **Technion – Israel Institute of Technology – Haifa, IL,** `avigal@technion.ac.il`
3 **RWTH Aachen, DE,** `jarke@dbis.rwth-aachen.de`
4 **TU Dortmund, DE,** `jakob.rehof@cs.tu-dortmund.de`

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19391 "Data Ecosystems: Sovereign Data Exchange among Organizations". The goal of the seminar was to bring together people from different disciplines (also outside the computer science area), in order to identify (i) a set of research challenges for the future development of data ecosystems and a catalogue of major approaches relevant to the field and (ii) a set of developed use cases of particular interest to the further development of data ecosystems. Towards the objectives, the seminar included tutorials, invited talks, presentations of open problems, working groups. This report presents the most relevant findings and contributions.

Seminar September 22–27, 2019 – http://www.dagstuhl.de/19391
2012 ACM Subject Classification Information systems → Information integration, Information systems → Data exchange, Information systems → Ontologies, Computing methodologies → Machine learning, Computer systems organization → Architectures, Security and Privacy, Information systems → Data analytics, Software and its engineering, Social and professional topics → Computing and business
Keywords and phrases Data sovereignty, Data ecosystems, Business models, Data integration, Ethics
Digital Object Identifier 10.4230/DagRep.9.9.66
Edited in cooperation with Bernadette Farias Lóscio (Collector)

1 Executive Summary

Cinzia Cappiello (Politecnico di Milano, IT)
Avigdor Gal (Technion – Haifa, IL)
Matthias Jarke (RWTH Aachen, DE)
Jakob Rehof (TU Dortmund, DE)

The design of *data ecosystems*, infrastructures for the secure and reliable data exchange among organizations, is considered as one of the key technological enablers for digitalization and the digital economy of the future. Several applied research initiatives and industry consortia provide substantive evidence of this trend e.g., the Industrial Internet Consortium (IIC)[1]

[1] https://www.iiconsortium.org/

formed in the USA, the Industrial Data Space (IDS) founded in Germany and the associated consortium International Data Space Association (IDSA)[2]. Most of these initiatives aim to provide a *reference architecture* for dealing with (i) *governance* aspects related to the definition of policies and conditions able to norm the participation to the data ecosystem, (ii) *security* aspects related to the definition of policies and infrastructures for guaranteeing a trusted and secure exchange of data, (iii) *data and service management* aspects related to representation models and exchange formats and protocols, and (iv) *software design* principles related to the realization of the architectural components and their interaction.

All these aspects have been discussed in the seminar and the main findings are described in this report. In addition, a central new aspect of data ecosystems that we considered in the seminar lies in the view of data as having an economic value next to its intrinsic value to support operational and decisional core business activities. This means that in the data ecosystem, data is typically considered both a business asset and a business commodity which may be priced and sold in some form (e.g., data provisioning service or raw data) according to contracts.

As testified by the amount and variety of problems described above, the creation of such ecosystems poses many challenges cutting across a wide range of technological and scientific specializations. For this reason, the seminar involved researchers from different communities. Interdisciplinary discussions gave the possibility to analyze different perspectives and to achieve valuable outcomes presented in this report, such as a wide set of research challenges and the definition of interesting use cases for the further development of data ecosystems. Details about the activities carried out during the seminar are provided in the following.

Overview of the activities

The seminar took place from Monday September 23 until Friday September 27. The seminar program encompassed four invited talks (keynotes and tutorials) on the first day (Sep. 23rd), by Gerald Spindler (law and ethics), Frank Piller (ecosystems and business models), Maurizio Lenzerini (data integration), and Boris Otto (International Data Space). After discussions related to the talks and tutorials, the remaining afternoon was spent structuring (through joint discussion) the coming days of the seminar and group structure. As a result, group structure was based on a thematic structure encompassing three groups, one for each of the topic areas Business, Data, and Systems. Tuesday Sept. 24 began with a breakout into groups and election of scribes in each of the three groups (Business, Data, and Systems), and the remainder of the day was taken up by parallel group sessions in the three groups. Wednesday Sept. 25 began with a joint session where each of the groups presented their work, which was then discussed jointly. The afternoon (until the excursion) was taken up by joint discussion on report structure. The morning of Thursday Sept. 26 encompassed joint discussion on a proposed joint manifesto as well as group discussions on application domains and application scenarios (topic areas were Health, SmartCities, Industry 4.0). The afternoon was taken up by continued group discussions and ended with group presentations and joint discussion on application domains and application scenarios. There was also further discussion on report structure at the end of the day. The manifesto was subject to very lively discussion in the evening, after dinner. Friday Sept. 27, the last day of the seminar, was

[2] https://www.internationaldataspaces.org/

devoted to wrap-up (conclusions, summary, and report process) followed by joint discussion on relations between Systems, Data and Business views on the overall topic of the seminar.

The outcome of the seminar, which is documented in the remainder of this report, encompasses summaries of the group discussions and the joint manifesto.

2 Table of Contents

Executive Summary

Overview of Invited Plenary Talks

Technology

Use Cases

3 Overview of Invited Plenary Talks

3.1 Recent developments of a legal framework for IT

Gerald Spindler (University of Göttingen, DE)

In his highly stimulating and provocative keynote address, law professor and high-level EU advisor Gerald Spindler shared important observations about a serious misfit between the conceptualizations pursued in business administration and computer science, and the structuring of the law system. As a consequence, judges are often surprised by seemingly unpredictable and contradictory answers to their legal questions. Conversely, managers and engineers are confronted with a law system that seems to adapt extremely slowly to the rapid progress, and with for them very surprising interpretation of this changing reality. For example, the speaker surprised the audience with the statement, that no concept of data ownership exists in Europe, except for the right to ones own personal data in the GDPR regulation. In the discussion, all participants agreed that joint research is urgently needed to better match the conceptual world of law and ethics, with the technical, user, and business perspectives on data ecosystems.

3.2 A few thoughts about managing business models for platform-based data ecosystems

Frank Piller (RWTH Aachen University, DE & MIT Media Lab, US)

In his keynote, Frank Piller started from the observation that – in contrast to traditional product platforms e.g., in the automotive industry – the value of smart products does no longer lie in the product itself, but rather in the connections in its business ecosystem. The resulting network effects and multi-sided markets have already been intensely studied, most visibly by 2015 Economics Nobel Prize winner Jean Tirole. The strategic question then becomes whether a company wants to offer a platform or an app. Successful platform businesses show striking differences from traditional organizations e.g., in terms of value created vs. value captured per employee. Current empirical research at RWTH Aachen University is studying how these concepts are being transferred to networks of German small and medium enterprises, e.g., in the smart farming sector.

3.3 Semantic Data Interoperability

Maurizio Lenzerini (Sapienza University of Rome, IT)

In his tutorial, Maurizio Lenzerini gave an logic-based structuring of four different data interoperability architectures: data integration, data exchange, data warehouses/data lakes, and collaborative data sharing. He pointed out the central importance of formal mappings with different schemas, and then went into some depth into description logic-based approaches for compile time tasks such as mapping discovery, analysis, and reasoning, but also into runtime tasks such as data exchange, update propagation, data quality, and mapping-based direct and reverse query rewriting.

3.4 International Data Spaces

Boris Otto (Fraunhofer ISST & TU Dortmund, DE)

In his overview talk on the Fraunhofer-led International Data Spaces (IDS) initiative, Boris Otto first presented some typical examples from European industries, and cited worries of a majority of European SMEs about trust and security, data sovereignty in terms of keeping control over shared data, and inconsistencies in not just data interoperability, but also process interoperability. He then reported a large-scale, multi-year requirements study about desirable properties of a solution, from which the IDS reference architecture as well as a conceptual information model, several advanced algorithms (e.g., for data usage control) and suitable governance mechanisms for so-called alliance-driven platforms are emerging.

4 Technology

4.1 Systems

Boris Düdder (Univerity of Copenhagen, DK) and Wolfgang Maaß (Saarland University, DFKI, DE) and Julian Schütte (Fraunhofer AISEC, DE)

4.1.1 Motivation

Data ecosystems operating valuable data assets need strong guarantees on data security while fostering openness, community, and value creation.

Data is becoming a significant asset for any organization [23]. A key challenge is related to technical architectures for managing and processing data. A dominant approach centralizes data and distributes processing results (ref). Alternatives are less popular but feasible, such as peer-to-peer architectures or other more distributed approaches (ref) in which either data or program (query) are transferred optimizing for connection capacities. Related to the underlying architecture are organizational principles and governance structures. For centralized architectures, principal-agent logic is applied, i.e., a principal (user, client, customer) sends data to an agent (server, service provider) who processes the data and sends back results. This organizational architecture is applicable if the key value lies in the software as such, but not in the data. Now that data become a value as such, a bilateral principal-agent, or technically a client-service architecture, does not necessarily fulfill the requirements of data economic systems.

Several technical elements are essential key components for a data ecosystem, as explained in the following subsections, such as security and encryption, data analytics, AI and semantic technologies and ontologies, multi-agent systems, peer-to-peer architectures and, last but not least, algorithmic correctness and correctness of implementations.

4.1.2 Scope and Requirements

The system boundaries are used to define the purpose of our data ecological system w.r.t. technologies and applications. Therefore, these system boundaries influence fundamental design decisions of the problem model design and are reflected in the solution design. The boundaries are defined along the dimensions of the organization and technical scope. Organizational boundaries are organized along with entities' roles, rights, obligations, and prohibitions. Such entities are key concepts such as contracts specifying obligations, data/data objects/data sets representing valuable curated data assets, and semantics/metadata/ontologies generating information out of pure data. The technical scope should support system design and modules, various pervasive architectures, i.e., central, tightly coupled, and loosely coupled, needs for adaptability [12] and security, as well as protocols, computation and programming models.

Data formats and protocols are necessary for expressing contracts, i.e., for enforcing policies for data usage and conditions or guards such as pricing or billing for example. A challenge in a distributed network to enforce contracts and usage constraints at a remote peer, which is controlled by another participant. Possible ways for enforcement are by establishing trust in the participants and their data handling, e.g., employing remote attestation [26], or by private function evaluations [21].

4.1.3 State of the Art

Traditionally data is not perceived as a valuable asset as such but as an intrinsic part of any software. Data populates databases as a carrier of facts about a domain. Physicists, chemists, astronomers, biomedicial researchers, and economists understand for a long time that data is a basic asset that requires processing and filtering for deriving knowledge (e.g., CERN LHC). Companies, such as Google, Amazon, Microsoft, Baidu, and Alibaba, adopted this understanding for extracting knowledge that can be used for predicting human behavior. The success of these attempts was transferred to production industries (Industrie 4.0 and cyber-physical systems) and even private life (smart city and smart home). Thus, data becomes a valuable asset as such, called a data product [62]. Taking a product-driven approach, system designs start with the data product and ask in which market, to whom, and for what price it can be sold, i.e., applying a *product logic*. This adopts an inside-out perspective by which a data product is the starting point, and the market is the target. From an inside-out perspective, prices for data products are determined by adopting a cost-driven approach. Taking an outside-in perspective, the customer of a data product is the starting point by asking the question of which value can be created on the customer side by leveraging data products. This generally applies a *service logic*.

Data ecosystems are characterized by the type of business and the type of community [79]. The terms 'knowledge market' and 'data market' are used as synonyms nowadays. The type of business is distinguished between commercial and non-commercial data ecosystems while the type of community distinguishes between closed on open. Closed data ecosystems are established under the umbrella of one juridical person. This could be, for instance, a corporation, an association, or an individual person. Non-commercial, closed data ecosystems are intraorganizational while non-commercial, open data ecosystems provide data products unlimited for free or with the option for donations, e.g., UC Irvine Machine Learning Repository. Viable business models for data ecosystems are closed, for instance, by membership or proprietary bi- or multi-lateral contracts. Commercial, open data ecosystems are organized as marketplace [24] that support trading and transactions [4].

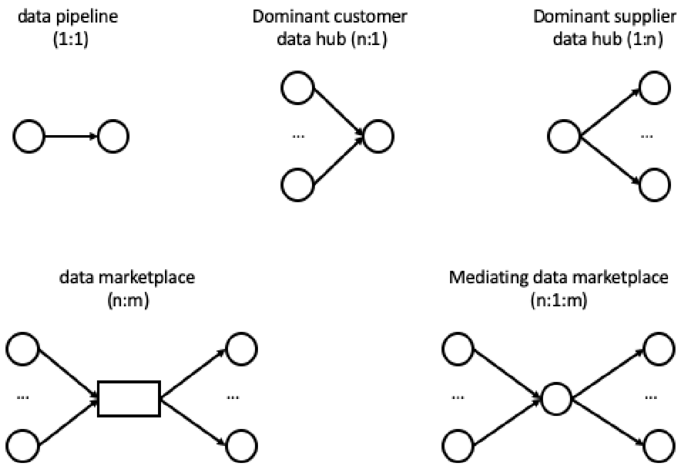


Figure 1 Basic system design patterns for data ecosystems.

System designs of data ecosystems taking a product logic adopt the concept of a marketplace in which the main interaction occurs by matching data product providers with customers. Facilitating roles are providing support services on financial transactions, logistics, quality assurance, insurance, and notary services. The main purpose of a product-driven marketplace lies in the transfer of access and usage rights under governance of explicit contracts or background regulations and laws.[3] The most simple system design is a *data pipeline* that establishes a marketplace between a single data provider and a single customer. Facilitating services are fully materialized by the marketplace and fixed contracts. More flexible are *data hubs* that allow $n : m$ transactions between data providers and customers. A $n : 1$ *data hub* is used if there is only one customer but many data providers. Dominant customers establish data hubs and invite data providers (*dominant customer data hub*). In contrast, a $1 : n$ *data hub* has only one or very few data providers but many customers (e.g., Airbus Skywise or Bloomberg Market Data Feed). A $1 : n$ data hub system design implements a standard product distribution network (*dominant supplier data hub*). A derivative of this system design is a *1:1:n data hub* with an independent agent implementing a data hub between one data provider and n customers. In reverse, the same system design can be applied with one dominant customer, i.e., $n : 1 : 1$ *data hub*.

More elaborated are $n : m$ *data hubs* or $n : 1 : m$ *data hubs* that constitute the concept of a *data marketplace*. Without dominant market players, a data marketplace is governed by offer and demand [80]. A data marketplace is considered being *liquid* if any demand can be contractually satisfied by an offer for a mutually acceptable price and low transaction costs. In plain data marketplace, trust between buyers and sellers is not guaranteed but requires additional trust-building regulations and additional roles.

The support for trusted and secure computation balancing participants requirements and expectations is challenging. Electronic marketplaces with multiple agents have been studied under various security objectives, including privacy, non-refutation, and accountability. The focus has primarily been on the digital transactions of such markets and how to secure them. The physical value-chain is aligned to the digital transactions of such marketplaces because

[3] According to Gerald Spindler's statements, data ownership cannot be transferred from a data originator to any other agent.

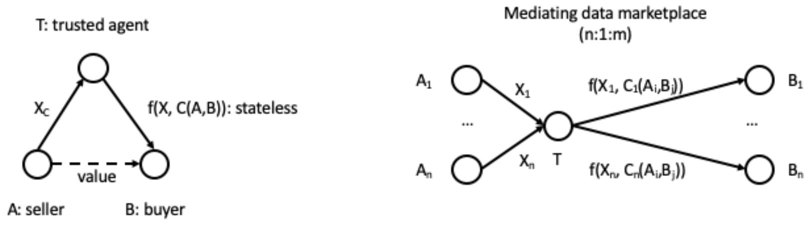


Figure 2 Generic system design for data marketplaces.

these marketplaces emphasize on physical goods as transaction value and not on data as a valuable good. To secure consistency and correctness of transactions, at least two different approaches are in use. The first approach employs monitoring and auditing of the processes, checking the consistency of good transfers and digital representation, as well as ensuring that data flows adhere to contractual agreements. The costly and laborious approach motivated due diligence [36] and various usage control frameworks [99, 92, 100]. Another approach is prevention by either using legal penalties to force compliant behavior of market participants or by technical measures including cybersecurity and cryptography. The latter has become a highly active area of research and several significant advances have been made in privacy-enhancing technologies such as verifiable computation and non-interactive zero-knowledge proofs [8, 1, 112, 18, 102, 7], distributed ledger technology [16, 17], and fully-homomorphic cryptography [50, 73]. Depending on the desired guarantees for the system, the correctness and robustness of algorithms and the system implementations are necessary. Security is not a product; it is a process, which demands a verified tool-chain, e.g., akin CompCert [77] and platform, i.e., trusted execution environments. The interaction of components and their reliability could be supported by automatic program generation [56, 11], DSLs [39] and verification methods.

4.1.4 Model

Data pipelines and data hubs replicate traditional interactions in business relationships between firms. Data marketplaces are a model class subsuming all other types of data ecosystems. Three roles are key to a data marketplace: buyer, seller, intermediary trusted agent. The role of a trusted agent is mandatory because of the transactional nature of a marketplace that requires a transaction being atomic. Data products X are provided from seller A to buyer B in exchange to another entity, in particular, financial assets. By doing this, value is created on the buyer side. In general, a seller A only transfers projection of X to buyer B for various reasons, such as IP protection and trade secrets. Nonetheless, a seller A might have a business incentive for selling certain properties and insights on X. Therefore, seller A and buyer B negotiate a function f that shall be allowed to apply on X under contractual requirements C. The role of the trusted agent T is that only f is applied to X as specified in C (Figure 2).

Expression $f(X, C(A, B)) : \mathit{stateless}$ means that a function f is applied to a data product X or a set of data products according to specification given by contract C acting as a callback function in f, e.g., in [78]. The result is provided to buyer B according to specifications in contract C. Function f is required to be implemented in a stateless fashion, and no traces will be left with agent T. These needs to be assured by trust-building measures, such as program verification and certification. Therefore, a system design consisting of seller A, buyer B, and a trusted agent T guarantees that a) data products are only used according to a specification, b) buyer B only gains access to resulting data, and c) no data traces are left with agent T.

4.1.5 High-level Description of Possible Applications

We present here two use case scenarios which could benefit from such a system.

Predictive maintenance – Predictive maintenance helps to determine the condition of operational equipment in-use for predicting when maintenance should be performed and which components are affected. Cost savings over routine or time-based preventive maintenance could be achieved because activities are triggered only when necessary. In this scenario, a production site creates raw data from production machine sensors, and the machine manufacturer is interested in running analytics on the sensor data to improve his customer maintenance service. On the other hand, the production plant owner is concerned about sharing raw data of the production machines, which could allow interested parties, e.g., competitors, to infer business secrets. The scenario can be enabled with a secure data sharing which balances the interests of both parties or even third parties.

Pharmaceutical companies – Pharmaceutical companies are producing and delivering drugs to patients through complex supply chains. The associated indirection between pharmaceutical company and patient complicates the due diligence, e.g., in the presence of adverse effects of drugs. Thus, sharing information between pharmaceutical companies, regulators, health practitioners, and patients could improve drug safety and reduce the costs of due diligence. All parties in this scenario have interests to share data but, at the same time, have high interests in restricting the data exchange, e.g., for privacy or competitive advantages. The system is realizable using secure data sharing allowing to create a data ecosystem harmonizing the needs of all participants.

4.1.6 Research Challenges

The project is ambitious, and to our best knowledge, there is no product on the market dealing with the identified challenges on business, legal, and computer science side. Business is founded on trust and trust-building activities, which is expensive to build and to maintain. The research challenges identified are (semi-)automatic contracts negotiation and agreement on functions as well as function execution. Semantic information of data and functions is necessary to automate negotiation and agreement on functions. On the other hand, data-/function-specific guarantees of market participants need to comply with data protection rules. The treatment of data in law is very often focused on privacy but not on data as an asset. The consequent juridical uncertainty, e.g., on the data ownership and copyright, is a business risk and poses a challenge for law researchers as well as legislation. To replace trust with mathematical and technological guarantees is an additional research challenge for computer science as a discipline. The verification of functions, their execution in secure execution environments, and secure and trusted libraries are not available on the market and require research from various interdisciplinary and inter-subdisciplinary research. The system challenges are compelling but indispensable for the future success of data ecosystems.

4.1.7 Conclusion

Because of the interdisciplinary aspects of data ecosystem platforms, computer science can provide various models, methods, and tools to support the development of such platforms.

The challenges are not only located in computer science but also the interplay with other disciplines. The challenges and opportunities presented in this section are novel in combination with their usage domain and its inherent restrictions, e.g., legal or economic. We postulate a platform and sound formal models supporting the activities described in this report and offering solutions to the more general problems of data ecosystems.

4.2 Data

Cinzia Cappiello (Polytechnic University of Milan), Yuri Demchenko (University of Amsterdam, NL), Ugo de'Liguoro (University of Turin, IT), Bernadette Farias Lóscio (Federal University of Pernambuco, BR), Avigdor Gal (Technion – Haifa, IL), Sandra Geisler (Fraunhofer FIT – Sankt Augustin, DE), Maurizio Lenzerini(Sapienza University of Rome, IT), Paolo Missier (Newcastle University, GB), Barbara Pernici (Polytechnic University of Milan, IT), Jacob Rehof (TU Dortmund, DE), Simon Scerri (Fraunhofer IAIS – Sankt Augustin, DE), Maria-Esther Vidal (TIB – Hannover, DE)

4.2.1 Scope

In this section, the main outcomes of the group of *Data* are reported. First, diverse data-driven problems are presented; solutions to these problems demand to effectively describe and integrate heterogeneous data, accurately represent and assess data quality, and ensure data specifications during the execution of operators or queries. Next, the state of the art is summarized and the proposed model for data ecosystems is defined. Finally, insights of the research challenges to be addressed are outlined.

4.2.2 Motivation and Requirements

Data-driven technologies in conjunction with smart infrastructures for management and analytics, increasingly offer huge opportunities for improving quality of life and industrial competitiveness. However, data has grown exponentially in the last decades as a result of the advances in the technologies for data generation and ingestion. Moreover, data is usually ingested in myriad unstructured formats and may suffer reduced quality due to biases, ambiguities, and noise. Thus, the development of efficient data management methods are demanded for enabling the transformation of disparate data into knowledge from which actions can be taken in scientific and industrial domains. Problems to be addressed include:

- Definition of a generic data market architecture that can describe data properties such as economic goods and data exchange models. In this type of architectures, data lakes play a relevant role in the storage of huge amount of heterogeneous data on distributed infrastructures.
- Active data networks described in terms of specifications and collections of components which can be repositories of data or queries. Data sources are connected using schema mappings. The evaluation of queries against the network requires the composition of data sources respecting the restrictions imposed by the connections existing the components of the network.

- Accurate description of meta-data of data sources that can change over time and effective usage of these descriptions whenever data is shared.
- Models that estimate the cost of integrating different sources, and the benefits that the fusion of new data sources adds to the accuracy of query processing.
- Hybrid approaches that combine computational methods with human knowledge; limitations and benefits of these approaches in the resolution of data-driven problems, e.g., schema matching, data curation, and data integration need to be established.
- Paradigms for making data suitable for sharing are required; they should ensure trustworthiness, e.g., in terms of data quality or data market players.
- Federated query processing for addressing data interoperability issues during query execution demands meta-data to select the relevant sources for a query. Furthermore, data quality assessment is required for determining the quality of the answers produced during query execution against a set of selected sources.
- Management of new generated data from data analysis and machine learning performed on existing sources.

4.2.3 State of the Art

4.2.4 Data Integration

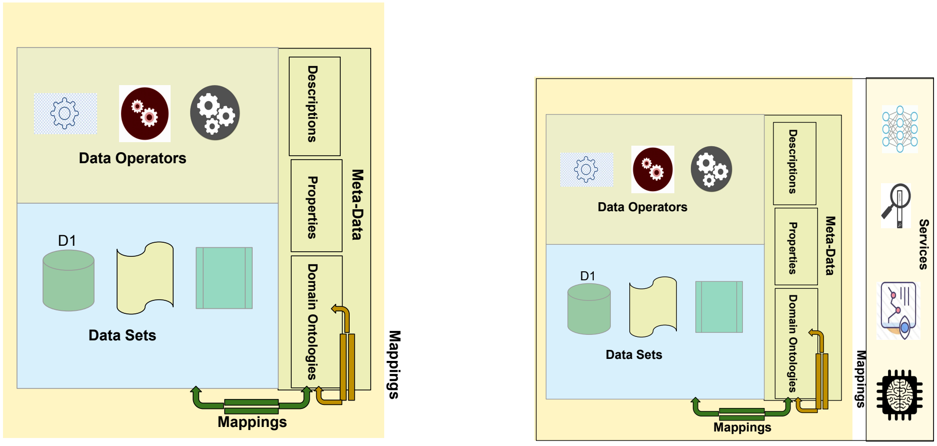
The problem of integrating data collected from different data sources has been extensively treated in the literature [34, 55]. The mediator and wrapper architecture proposed by Wiederhold [114] and the data integration system approach presented by Lenzerini [74], represent the basis for the state of the art [27, 52, 65, 83]. Following these approaches a global ontology encodes domain knowledge and enables the description of the meaning of the data to be integrated by means of mapping rules. Since creating mappings manually in a data integrating system is a tedious and time consuming task, diverse approaches have been proposed to discover mappings in a (semi-)automated way (e.g., CLIO[42], IncMap [94], GeRoMe [61, 54], KARMA [65]). Additionally, a vast amount of research has been conducted to propose effective and efficient approaches for ontology alignment or schema matches [22, 41, 45, 81, 85]; modeling and management of uncertainty of the schema matching process is the paramount importance for providing a quantifiable analysis of the accuracy of the discovered mappings [44].

4.2.5 Data Ecosystems

Data ecosystems are a special kind of digital ecosystem. As such they are distributed, open and adaptive systems with the characteristics of being self-organizing, scalable and sustainable. While being centered on data, the main concern with data ecosystems is about knowledge sharing and growing, which is at the same time an issue of learning from unstructured and heterogeneous data, construction of new abstractions and mappings, offer of services, including querying, data integration and transformation. All this should be ensured in a dynamic and scalable way, while retaining consistency, quality assessment, security and affordability.

4.2.6 Query Processing Over Heterogeneous Data Sets

Existing solutions to the problem of query processing over heterogeneous data sets rely on a unified interface for overcoming interoperability issues, usually based on metamodels [63]. A few Data Lake systems have been proposed, mainly with focus on data ingestion and metadata extraction and management. Exemplary approaches include GEMMS [96],



(a) A Basic Model for a Data Ecosystem. **(b)** A Data Ecosystem equipped with a brain.

Figure 3 A Basic Model for a Data Ecosystem.

PolyWeb [64], BigDAWG [37], Ontario [40], and Constance [54]. These systems collect meta-data about the main characteristics of the heterogeneous data sets in a Data Lake, e.g., formats and query capabilities; additionally, they resort to a global ontology to describe contextual information and relationships among data sets. Rich descriptions of the properties and capabilities of the data have shown to be crucial for enabling these systems to effectively perform query processing.

4.2.7 A Model for Data Ecosystems

A data ecosystem *DE* is defined as a 4-triple *DE=<Data Sets, Data Operators, Meta-Data, Mappings>*; Figure 3a depicts a data ecosystem in terms of its components.

- *Data sets*: the ecosystem is composed of a set of data sets. Data sets can be structured or unstructured; also, they have different formats, e.g., CSV, JSON or tabular relations, and can be managed using different management systems.
- *Data Operators*: the set of operators that can be executed against the data sets.
- *Meta-Data*: provides the description of domain of knowledge, i.e., the meaning of the data stored in the data sets of the data ecosystem. It comprises:
  i) *Domain ontology* provides a unified view of the concepts, relationships, and constraints of the domain of knowledge. It associates formal elements from the domain ontology to each D. For instance, *workshop* and *participant* can be part of the concepts in a domain ontology.
  ii) *Properties* enable the definition of data quality, provenance, and data access regulations of the data in the ecosystem. For instance, *last updated* and other non-domain properties (quality etc).
  iii) *Descriptions* of the main characteristics of a data set. No specific formal language or vocabulary is required; in fact, a data set could be described using natural language. For instance, *Data set D is about a Dagstuhl seminar.*
- *Mappings* expressing correspondences among the different components of a data ecosystem. The mappings are as follows:

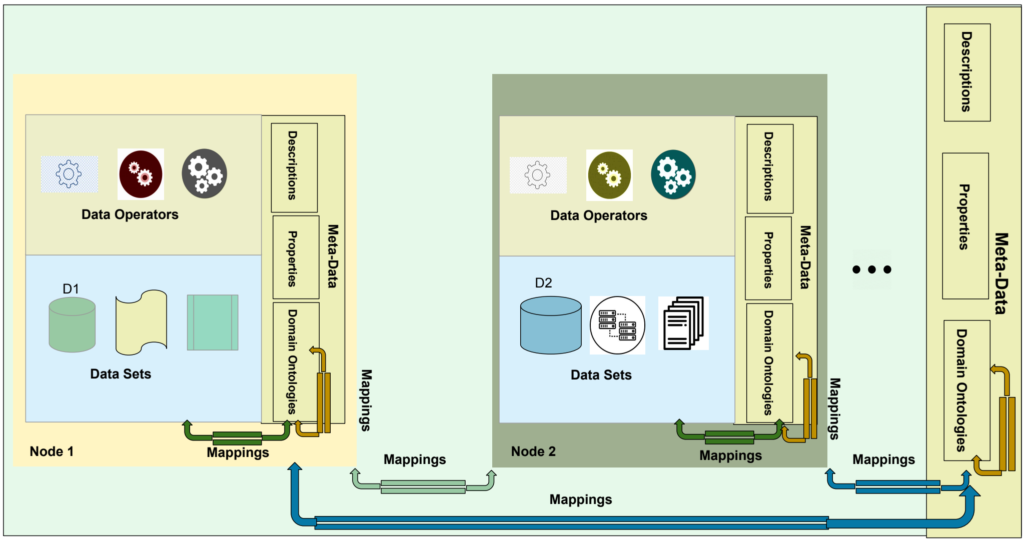


Figure 4 A Model for a Network of Data Ecosystems.

- *Mappings between ontologies*: they represent associations between the concepts in the different ontologies that compose the domain ontology of the ecosystem.
- *Mappings between the data sets*: they represent relations among the data in the data sets of the ecosystem and the domain ontology.

A data ecosystem can be equipped with a "brain", able to execute services against the data sets (Figure 3b). Services include query processing, data transformation, anonymization, data quality assessment, or mapping generation. The services are able to exploit the knowledge encoded in the meta-data and operators to satisfy the requirements of the applications implemented across the ecosystem. The following correspond to examples of services:

- *Concept discovery*: identify a new concept, e.g., *foreign student*, using machine learning. Based on the result, the domain ontology and the mappings can be augmented.
- *Data set curation*: a service able to keep humans in the loop can be used to create a curated version of a data set in the ecosystem. The service can also update the properties of the ecosystem to indicate the provenance of the new curated data set and manage new generated data from data transformation, analysis, and learning.
- *Procedure synthesis*: a service able to construct new procedures out of elementary build blocks by composing existing services toward new goals. In a complex and evolving system, it would be unfeasible to program procedures and even queries without automatic support; also, the exploration of repositories and libraries of existing procedures should be available.

A set of data ecosystems can be connected in a network. For this, we envision an *ecosystem-wide* meta-data layer where the entire ecosystem is described. Figure 4 depicts a network where nodes and edges correspond to data ecosystems and mappings among them, respectively. In this configuration, the meta-data layer describes each of the nodes in terms of descriptions, properties, and domain ontologies. The following types of mappings can be defined among the nodes of a network of data ecosystems:

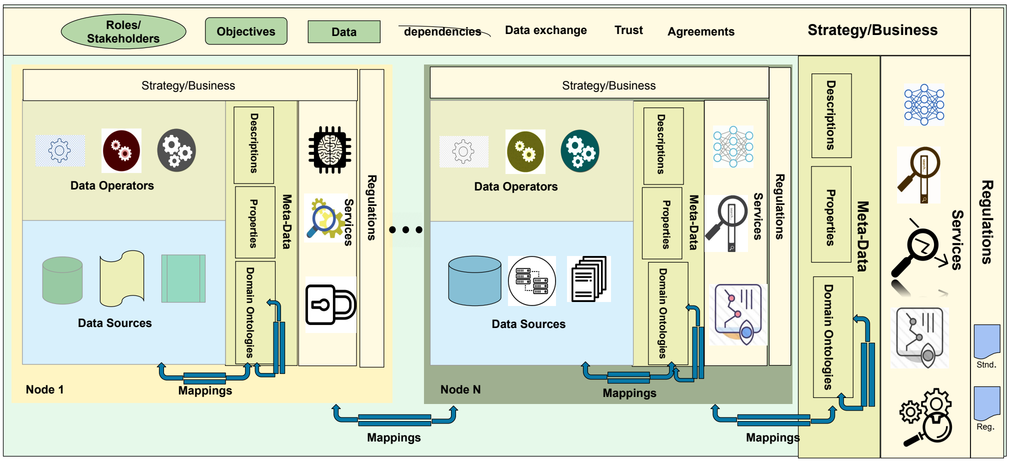


Figure 5 A Model for a Network of Data Ecosystems Empowered with Strategy and Business Models and Regulations.

- *Mappings between domain ontologies*: they state correspondences among the domain ontologies of two nodes or between one node and the global meta-data layer. For example, the concepts *workshop* and *seminar* in nodes N_1 and N_2, respectively, are the same.
- *Mappings between properties*: they describe relationships among properties in two nodes. For example, the provenance of two curated versions of a data set could be the same.
- *Mappings between data sets*: they represent correspondences between the data sets in two nodes. For instance, mapping data from D_1 in node N_1 to D_2 in node N_2, can represent that the list of students in a university 1 is the same to list of students in university 2.

Finally, data ecosystems can be enhanced with additional meta-data layers to enable the description of business strategies and the access regulations. Figure 5 depicts the main components of a network of data ecosystems empowered with these layers. As can be observed, this enriched version of a network of data ecosystems comprises:

1. Meta-data describing business strategies will enable the definition of the stakeholders of the network and their roles.
2. Objectives to be met and the dependencies among the tasks that need to be performed to achieve these objectives.
3. Agreements for data exchange and criteria for trustworthiness.
4. Regulations for data access and for data privacy preservation.
5. Services composing services of the nodes of the network or exploiting their operators.

4.2.8 Research Challenges

This section presents an outlook on the main challenges to be addressed in the implementation a network of data ecosystems.

4.2.9 Data and Metadata Curation

The availability and quality of data resources must be ensured, so that data value creation can be stimulated. A promising solution is to use a well-conceived, efficient curation strategy for data resources and their metadata. Such a curation technique is the continuous process of managing, improving, and enhancing the data and their metadata. Furthermore, the curation process aims to ensure that the data and metadata meet a defined set of quality requirements, such as security rules, integrity constraints, or metadata availability expectations. Without proper curation, data resources may deteriorate in terms of their quality and integrity over time. One of the major challenges for achieving continuous curation of metadata is to create a methodology to structure the curation process as well as to provide a set of tools. Furthermore, data sets obtained through transformation processes of data analytics or machine learning can become new data sets. Not only provenance but also the processing methods should be associated with them.

4.2.10 Data Traceability and Data Consumption Monitoring

Tracking the utilization of data sets and applications using these data resources is still a big challenge. Such information could be useful for both the identification of new data sets and for data quality improvement. Monitoring data consumption, as well as providing effective ways for the consumer to interact with the data publisher, should make it possible to collect information about using and sharing data. In this sense, it is crucial to obtain consumers' feedback in such a structured way that it allows identifying flaws in the published data, the need to publish new data, and, for example, to enable classification of data.

4.2.11 Describing the Main Properties of Data Sets

All data sets in a data ecosystem should be described in terms of their main characteristics. In particular, in the case of numerical data, a general framework that allows for the representation of uncertainty and imprecision is required. In addition, keeping track of data set updates and modifications, and allowing basic access to versioning information are also important issues. For instance, a new version may be created when there is a change in the structure, contents, or characteristics of a data set. As data sets can change over time, maintaining different versions of the same data set and enable access to them becomes necessary.

4.2.12 Data Enrichment via a Joint Human/Machine Integration Effort

Data integration in a data ecosystem is challenged by the need to handle large volumes of data, arriving at high velocity from a variety of sources, which demonstrate varying levels of veracity. Data integration has been historically defined as a semi-automated task in which correspondences are generated by matching algorithms and subsequently validated by a single human expert. The reason for that is the inherent assumption that *humans do it better* which is not necessarily the case. A current challenge involves the identification of respective roles of humans and machines in achieving cognitive tasks in a way that maximizes the quality of the integrated outcome. We observe that the traditional roles of humans and machines are subject to change due to the availability of data and advances in machine learning.

4.2.13 Adaptive/Context-aware Data Quality Assessment and Redefinition of Data Quality Dimensions

Data quality assessment requires the selection of both the dimensions to be evaluated and the metrics to measure the selected dimensions. An accurate evaluation of the quality of the data depends on diverse factors, e.g., the type of source or data, and the application that aims to use the data. This implies an adaptive approach for data quality assessment able to trigger the appropriate metric. Furthermore, data quality assessment should be performed in two phases: the registration and the usage phase. During the phase in which data are registered/ingested into a data ecosystem, a first evaluation is needed. This evaluation should consider some basic metrics in order to guarantee that low quality data is not ingested in a data ecosystem. On the other hand, in the usage phase, metrics that reflect the suitability of a data set along the applications that aim to access the data set need to be defined. Moreover, data quality can be measured at different levels. There are quality dimensions which refer to the schema or overall structure of a data source. Additionally, data quality dimensions regarding the content of the data sources can be defined. In both cases, metrics which represent functions to determine the corresponding values for the data quality dimensions, may cover atomic items, such as a single attribute, or multiple items, or complex objects. In a single data ecosystem node, but especially in the case of sharing data between several nodes, it is not clear, how data quality values should change, when the data is further processed. The way to calculate new values for a single dimension might be dependent on the meaning of the dimension and a given metric, as well as specifically on what happens to the data in the current processing step.

4.2.14 Measuring Information Gain in a Data Ecosystem

Adding new data sets to a data ecosystem arguably results in some kind of information gain. The gain extends naturally the increase in knowledge that occurs when new mappings are discovered within a data ecosystem node – either interactively or semi-automatically, or across diverse nodes in a data ecosystem. Questions that need to be addressed include:

i) How is the information gain defined and measured?
ii) Can the gain be defined in terms of benchmark queries and the results they return given an evolving state of a data ecosystem?

4.2.15 Exploring the Social Dimension of a Data Ecosystem

Users (humans, or data consuming services) are actors in a data ecosystem, playing an important and active role in its evolution. Firstly, they can provide feedback on data sets that they have used (either formally or informally), adding to the quality and provenance properties, which other users should be able to take into account. Secondly, they may decide to curate data sets that exhibit poor quality properties and return a better version of those to the DE, generating more metadata in the process (provenance tracing with details of the interventions). Finally, they may contribute to the mapping exercises that the DE makes possible and indeed relies upon. Some of the questions that emerge when users are first class citizens include:

i) Should data ecosystem users form a network, and if so, can such a network be used to recommend / promote / deprecate data sets?
ii) Should users (or services) be credited for producing and making available new (better) versions of a data set, i.e., through curation activities?
iii) How can the social layer of the DE be promoted, maintained (is that a part of the "upper layer"?), and exploited?
iv) How would a recommender system be able to suggest data sets to users based on their usage history of the data ecosystem and the implicit connections with other users?

4.2.16 Explainability in Data Ecosystems

Whenever the *brain* of a data ecosystem executes a service (e.g., query, data transformation, quality evaluation, or mapping generation), it should be able to explain the result of the operation carried out in executing the service using an appropriate language. There is a connection between provenance and explanation. In some sense, provenance information can be part of/exploited in an explanation. Included in the notion of explanation is the idea of explaining the semantics of a data source–either primitive, or produced by a data operation.

4.2.17 Query Processing over Data Ecosystems

Effectiveness and efficiency of query processing over a federation of data sources is known to be affected not only by the number of sources in the federation but also for the heterogeneity of these sources. In case of a Date Ecosystem, data sources can be ingested in different formats and accessible with management systems that provide different data management capabilities, e.g., some systems may support complex queries including joins and aggregations while others do not. Moreover, the data quality and structuredness may considerably vary. For ensuring the correct processing of queries, interoperability issues should be solved at different levels. Meta-data about the description of the meaning of the data stored in the data ecosystem plays a crucial role for enabling the execution of queries over the data sources that will provide the correct answers. Moreover, if several versions of the data sets are stored in a data ecosystem, the process of data integration must take into account the temporal dimension of data. For example, answering queries like *tell me all the students of University X who have become professors at X after no more than 10 years from graduation*, might involve looking at several versions of different data sets.

4.2.18 Data Ecosystems for enabling the specification of Meta-Metadata

The proposed model for Data Ecosystems enables the definition of a second layer of global meta-data which describes both data and meta-data descriptions of all the local nodes in the Ecosystem. Managing and keeping meta-metadata *up-to-date* is more challenging because both data and metadata may change as a consequence of the dynamics of such systems. Dynamicity may be at least twofold:

i) local dynamics of node data and metadata and
ii) global dynamics of queries and mappings involving several nodes and the Data Ecosystem metadata themselves.

The research questions to answer under these conditions are:

a) How to define safety criteria?
b) What kind of transformations should be allowed?
c) Should we prevent non-conservative structural updates?
d) Which level of data integrity and consistency will be ensured?

4.2.19 Specification language for procedure synthesis

Beside ontologies and meta-data, a language for specifying services and their functional properties is needed. This is not just for the sake of verification, but mainly because they are required for procedure synthesis. For automatic search through large libraries and for synthesis to be feasible such a language should balance expressive power with effectiveness and efficiency. Attempts to use type theory for these purpose include work by Bessai et al [10] and Hengelin and Rehof [58].

4.3 The Business of Data Ecosystems

Elda Paja (IT University of Copenhagen), Matthias Jarke (RWTH Aachen & Fraunhofer FIT), Boris Otto (Fraunhofer ISST & TU Dortmund) and Frank Piller (RWTH Aachen)

4.3.1 Background and Motivation

▶ **Example 1** (Data Ecosystem). *Consider the example of PRINT INC., an industrial print shop reacting to increasing demand by adding more flexible production capacity. The firm provides machine vendor DRUCKMASCHINEN AG a list of requirements, derived from its most sophisticated print jobs. DRUCKMASCHINEN, however, lacks insights about the specific situation and suggests a standard specification. After ramp-up, a sample analysis reveals an OEE far below the anticipated values, as configurations, parameters and procedures of the existing facilities do not recognize the capabilities of the new equipment. This scenario is still a very common situation in many industries. With a data ecosystem in place, PRINT INC. could first provide DRUCKMASCHINEN better past operating data to better determine requirements. A specialized analytics service provider would assist in this analysis, also drawing on technology insights and usage data from other print shops operating in a related setup. After implementation, a higher level of OEE could be reached by learning from best practices from similar operations of other manufacturers. By getting access to rich usage data – covering not just machine data, but also data on the production context, material characteristics and behavior of PRINT INC – OEE can be continuously improved. The data ecosystem enables a continuous feedback of this data into the development cycle of DRUCKMASCHINEN, revealing requirements for the next generation of production hardware and software. In times of low capacity utilization, PRINT INC. provides DRUCKMASCHINEN access to its new printing machine, which becomes part of DRUCKMASCHINEN's production network, connecting printing capacities from hundreds of printers virtually. With its PRINT HUB platform, DRUCKMASCHINEN can place print jobs from clients like publishing houses (without own production capacity) on existing machinery, generating a new business opportunity for both itself and its customer. At the same time, by moving from the role of a manufacturer and service provider of printing machine to that of an operator, DRUCKMASCHINEN also gains critical knowledge in the operating principles and continuous improvement opportunities of its machinery.*

To realize this typical vision for Industry 4.0, much more than a technical infrastructure is required. Beyond the availability of data and capacity, this scenario only can be realized if it has been determined how the value gained by accessing these inputs is captured and shared among the actors involved:

- Would profiting from such external data also require to provide similar feedback data?
- Could PRINT INC. control who and what is printing on its capacity shared via PRINT HUB?
- How could print shops differentiate themselves by process know-how when a data ecosystem enables a kind of instant benchmarking, as all operational data is shared over the network?
- Why is DRUCKMASCHINEN, in the end, not operating all its machines directly, integrating vertically into the domain of its former clients?
- Who owns the governance of this network? Is it DRUCKMASCHINEN as the focal keystone player, or should rather an alliance of print shops, users (like publishing houses) become the owner of the platform [113]?

This generic example provides a glimpse of the challenges from an economic and competitive dynamics perspective that come with the vision of a data ecosystem, which is, by definition, not restricted to a focal company or value creation within a closed network of established partners. It resembles the vision of an open network of sensors, assets, products, and actors that continuously generate data. This data is utilized to enhance operational efficiency, but also to provide new opportunities for strategic differentiation. A core element hence is a business-model perspective on **(re-)usage of data, insights, and applications by other parties** than those generating the data at the first place. Generally, learning and analytics can take place faster and more efficiently if manufacturers not only utilize their own data, but also could access data from similar contexts in other industries.

Actual research on industrial data-based business models has focused predominantly on the perspective of value creation, i.e. how to use shared data to create new service offerings like predictive maintenance, energy optimization, quality improvements, etc. [19]. But the open question still is an understanding how **to incentivize the sharing of deep production know-how and data** in order **to balance value creation (using the data) with value capture (sharing the rents)** [70]. The **rise of platforms** where these data are being exchanged and enhanced by dedicated "apps", often offered by specialized third-party entities, is **one of the largest economic developments of the last decade** [28, 91]. As platform **interfaces** become more open, more actors will be attracted into a data ecosystem, and the platform orchestrator will be able to access a larger set of potentially complementary innovative and operative capabilities. Most of the external contributors will innovate in ways complementary to the platform. Some, however, may start developing **capabilities** in ways that become competitive to the platform. Such an emergence of competition will depend on the **governance of the ecosystem**. Collaborative governance mechanisms will increase complementors' incentives to innovate in platform-enhancing ways. This demands especially a dedicated setup to share the value created on the original platform, but is also dependent from the technical interface design. In turn, emergence of competition from former complementors is likely to create a reaction by the platform leader to start competing back with these new rivals, either by enveloping them, or by closing its technological interface, in effect **moving away from being an open data ecosystem** towards becoming a more closely managed network or internal platform. Based on a discussion of different governance frameworks for data sharing and access, this workgroup of the **Dagstuhl Seminar #19391** discussed various scenarios, but especially questions for further research. Considering actual developments in platform-based business models that recently emerged also in manufacturing industries, our workgroup discussion tried to understand the design of the business models, but especially how to incentivize (reward) the sharing of deep production know-how and data in order to balance value creation (using the data) with value capture. This led to the following questions suggested for future research:

- Modeling the tension between openness in value creation and control of value capture, while recognizing the need for establishing and increasing trust in data sharing.
- Managing property rights (access, transfer, enforcement) at data, applications, and connected assets as a result of varying degrees of platform openness
- Definition of governance modes and design factors to generate adequate business models for a data ecosystem that allow to maximize value appropriation for all involved actors

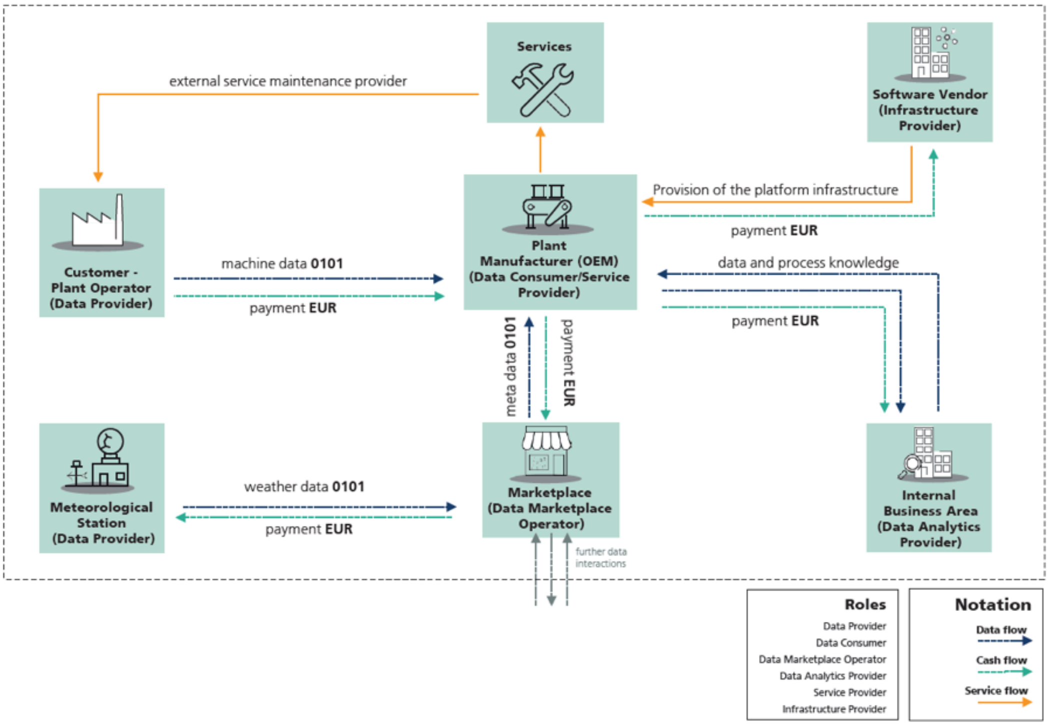

Services
Software Vendor (Infrastructure Provider)
external service maintenance provider
Customer - Plant Operator (Data Provider)
Plant Manufacturer (OEM) (Data Consumer/Service Provider)
Provision of the platform infrastructure
payment EUR
machine data 0101
payment EUR
data and process knowledge
payment EUR
meta data 0101
payment EUR
Meteorological Station (Data Provider)
weather data 0101
payment EUR
Marketplace (Data Marketplace Operator)
further data interactions
Internal Business Area (Data Analytics Provider)
Roles
Data Provider
Data Consumer
Data Marketplace Operator
Data Analytics Provider
Service Provider
Infrastructure Provider
Notation
Data flow
Cash flow
Service flow

Figure 6 Digital Business Ecosystem Example.

4.3.2 From Data Ecosystems to Business Ecosystems

Figure 6 illustrates the interactions within a digital business ecosystem in terms of data, cash and payment flow. Within the ecosystem perspective, a holistic view on the business relations is provided.

The digital ecosystem is enabled via multiple technical infrastructures. In this example two major infrastructure providers (data marketplace operator and software vendors provide a network to connect various entities such as the physical machines, customers, service staff, and third party services e.g. for data analyses. The Original Equipment Manufacturer (OEM) collects data through specially installed edge devices on the customer's sites. The operational machine data is transferred to the platform-based cloud infrastructure provided by the OEM. The data is stored, processed, and analyzed on the OEM's platform. Part of the data analysis is performed by the OEM itself within separate organizational units, so that the role of the data analytics provider is also taken internally. Therefore, additional historical and product development data is provided from internal resources to improve the analysis. Based on the analyses of the OEM, various services can be provided to the plant operators via the platform (e.g., dashboards on machine usage, specific operational reports, process optimization recommendations, maintenance planning). New external data sources (e.g., weather data) from meteorological stations can be considered as new parameters for analysis. This data is provided within a marketplace, which the OEM uses to enrich the existing data basis. The data marketplace acts as a gateway to other data ecosystems and promotes cross-industry data exchange. The ecosystem perspective provides an overview of the dynamics within a practical data ecosystem including the flow of data, payment and services.

4.3.3 State of the Art

One of the most important questions of a firm is how to efficiently create value: Should it produce its own output or should it orchestrate the output of others? In the case of software, the choice increasingly favors orchestration over production. Apple, Google, and Microsoft became the three most valuable companies in the world in 2015 by relying on a **platform business model** to provide software applications for its users, developed by independent app programmers. An open platform business model offers distinct economic advantages because it allows a firm to harness external inputs and innovation as a complement to internal innovation [91]. These platform markets are, however, neither a new phenomenon nor restricted to software or information goods. Open platforms have emerged in aerospace (Lockheed Martin), T-shirts (Threadless), 3D printing (MakerBot), and shoes (Adidas).

Since the early 2000s, the industrial organization literature has begun to develop theory on platforms (also referred to as "two-sided markets", "multi-sided markets", or "multi-sided platforms" [97, 98]). Economists view platforms as special kinds of markets that play the role of facilitators of exchange between different types of users that could not otherwise transact with each other. Essential to most economic definitions of **multi-sided platforms** (MSPs) are the existence of "**network effects**" that arise between the "two sides" of the market [49]. As the value of the platform stems principally from the access of one side to the other side of the platform, the question of platform adoption becomes how to bring multiple sides on board. Such platforms typically reside upon a **layered digital infrastructure**, where lower-level layers (e.g., physical components, transmission layer) enable and support functionalities at higher, user-facing layers (e.g., operating systems layer, application layer) [109, 115]. To create value, ecosystems hence depend on complementary inputs made by loosely interconnected, yet independent stakeholders from varying levels of (technological) distance from the end consumer [3, 90].

MSPs[4] facilitate the establishment of business ecosystems, which are formed by the users interacting on the platform. Researchers have examined a number of case studies on MSPs in which a keystone firm owns and governs the platform [86, 104]; so these are "keystone-driven MSPs". However, the platform landscape is becoming more and more diverse, with other, more complex governance and ownership structures being observable in different domains. De Reuver et al. [30] have found that in many cases there is no single platform provider, but that the platform is jointly designed and "shaped" by multiple actors. In this context, Tiwana et al. [110] point to the importance of distinguishing platforms owned by a single firm from platforms characterized by some form of "shared ownership". Shared ownership materializes in multiple organizational forms, among them the alliance. Gawer [48], for example, identified some MSPs from the supply chain management domain that are "shared among firms that are part of a formal alliance". Such "alliance-driven MSPs", which are characterized by shared ownership and governance (e.g. a joint-venture company or industry association), as well as decentral platform governance models have been neglected by academic research so far [30].

Table 1 compares keystone-driven with alliance-driven MSP design.

In the case of an **industrial data ecosystem**, platform participants include the **orchestrator of the platform** (today either IT infrastructure providers like SAP or Software AG, or providers of automation or manufacturing equipment like Siemens, GE, Trumpf), operators of **production assets** (users in form of "factories"), which provide data, providers

[4] This paragraph is taken from Otto and Jarke (2019) [89]

Table 1 Juxtaposition of Keystone-Driven and Alliance-Driven MSP Design.

Theoretical Concept	Keystone-Driven	Alliance-Driven
Platform architecture	Architecture determined by goals of keystone firm	Architecture determined by shared interest of multiple owners (leading to decentral data storage, for example)
Platform boundary resource	Mainly technical boundary resources (APIs, SDKs etc.), supported by "social" boundary resources (e.g. training for developers) [14]	Data (IDS specific) as a boundary resource of "dual" nature, i.e. requiring both technical processing and functional use; many social boundary resources, such as working groups, task forces etc
Platform design	Core developed by platform owner, then extended by complementors	Consensus oriented design process with focus on "common denominator"
Platform ecosystem	1) Innovation, 2) Adoption, 3) Scaling [57]	1) Adoption, 2) Innovation, 3) Scaling
Ecosystem governance	Start with limited number of sides and limited options for interaction between them, then increase number of sides and options for interaction [105]	Start with complex ecosystem (i.e. multi-stakeholder setting), then reduce to core ecosystem and extend it later on depending on roll-out requirements
Regulatory instruments	Mainly pricing instruments, accompanied by non-pricing instruments	Dominated by non-pricing instruments; integration of pricing instruments scheduled for scaling phase; data governance

of applications analyzing this data and providing prescriptions and predictions (**app programmers**), and other asset operators that utilize insights from aggregated data to optimize their own production. In addition, also the goods being produced can become part of the platform in form of **connected ("smart") products**, providing a feedback loop of usage data, but also becoming the center of another platform around digital services complementing these products. The latter also refer to **end-users (customers)** as a final participant of the ecosystem.

An applied stream of research looked into the demand for business model innovation (BMI) in leu of Industrie 4.0 [19, 69]. Based on rather qualitative research approach, a demand for professionalizing the BMI process in established companies was identified [95]. Similarly, specific aspects of BMI for Industrie 4.0, i.e. specific patterns or components of I40 BMs, have been identified. By analyzing Industrie 4.0 characteristics for a firm's business model and the approaches of pioneering companies to instigate BMI, this research developed a methodology to both generate, model, systematically design, and evaluate BM alternatives triggered by Industrie 4.0 [93].

Another recent stream of literature complements the original economic analysis of platforms with a more managerial and design-orientated perspective. It looks into the distinct **governance and orchestration challenges** presented by established innovation ecosystems (e.g., [33, 113]). This literature has looked, for example, into prices, incentives, contracts, and network effects. Less work has addressed how prospective ecosystem stakeholders **commit resources** towards a de novo ecosystem creation effort and how they evolve a shared structure of interactions [28]. Yet, this is exactly the situation of many data ecosystems, where creation is not a simple endeavor of an app. Rather, the ecosystem value proposition depends on the concurrent availability of complementary inputs from varied, independent stakeholders.

In such a situation, a core decision of a platform operator is about how much **to open the platform** and when to absorb inputs (developments, apps, data) from the connected parties. This decision drives adoption and harness developers as an extension of the orchestrator's own production function [9, 15, 91]. [28] extend the analysis to contexts in which the establishment of a new business ecosystem cannot be reliably planned, as no clear value proposition to orchestrate the ecosystem exists ex-ante. In a piece of complementary research, [9] investigate the perspective of platform complementors (app programmers) and their decision to join a platform based on its openness. The authors develop and validate a platform **openness measurement instrument** that captures perceived platform openness.

Since 2014, our Fraunhofer Industrial Data Space initiative [60, 87] has focused on requirements and rather technological challenges of inter-organizational data exchange. This requires novel conceptual information modeling at multiple levels, and still significant research for the specialization to the case of production engineering. The ideas pursued at the boundary of CRD_A and this workstream build on significant prior research on the conceptual modeling, structure and evolution analysis, and data-based Social Network Analytics concerning strategic dependencies and trust among players [46, 47, 84].

4.3.4 Conceptual Background[5]

The proliferation of digital technologies and AI accelerates a shift in business models which is characterized by an increasing importance of data as a strategic resource. While traditional business models rest on tangible assets, data is the raw material not only for information and knowledge, but for innovative services and customer experiences. Besides the shift from tangible to "smart" products and from controlling the physical to orchestrating the data value chain, there is one additional fundamental change in the digitalized economy. Innovation increasingly takes place in ecosystems in which various members such as businesses, research organizations, intermediaries such as electronic marketplaces, governmental agencies, customers, and competitors band together to jointly achieve innovative value offerings.

Ecosystems are characterized by the fact that no one member is capable of creating innovation on its own, but the ecosystem as a whole needs to team up. In other words: Every individual member has to contribute in order to benefit. Ecosystems function in an equilibrium state of mutual benefits for all members.

In a data ecosystem, data is the strategic resource for the success of the whole system as it is understood as a stand-alone asset that will be exchanged and monetarized within the ecosystem. That offers the participating actors new growth opportunities through networking with other participants and acts as a driver for innovative services and customer experience.

Data sharing opens up new opportunities for progress and the formation of cooperations with other companies or actors from which every participant in the data ecosystem benefits. The various activities of the different members in a data ecosystem lead to a complete coverage of the data value chain. This includes the stages of data generation, curation, exchange, storage and analysis as well as the use of the resulting knowledge for comprehensive business decisions. Through the sustainable data exchange the participating actors are able to develop further and to operate value co-creation which leads to new digital value proposition.

[5] Cf. Otto et al. (forthcoming). [88]

4.3.5 Data as a Boundary Resource[6]

As far as the boundary resources are concerned, Henfridsson and Bygstad [57] argue that this concept is helpful for studying patterns of interaction between the various groups and agents on a digital platform. Boundary resources are resources through which different agents create relationships and interact with each other in order to co-create value [38]. Dal Bianco et al. [14] distinguish between technical and social platform boundary resources. Typical boundary resources are Application Programming Interfaces (APIs) and Software Development Kits (SDKs). Examples for social boundary resources are intellectual property rights and documentation of software services. Furthermore, boundary resources are not stable, but evolve over time. Eaton et al. [38] coin the notion of "distributed tuning" to describe the process of continuous shaping and reshaping of boundary resources between the different platform actors and users. More recent research has suggested to increasingly look at such boundary resources of digital platforms as a promising subject of analysis [30].

How organizations can exchange and share data has long since been an important research topic. The need for companies to exchange and share data has been a major motivation for the development of platforms mediating between suppliers and buyers of goods. Early two-sided data exchange solutions were facilitated by technological standards, such as EDIFACT or ANSI X.12. Gawer [49] within her integrative platform framework identified traditional buyer-supplier relationships for which data is a technical enabler.

Around the turn of the millennium, electronic marketplaces emerged as intermediaries to reduce the complexity of the increasing need of n:m data exchange [101], in which data from multiple sources (n) can be bundled and utilized in contextualized presentations to multiple users (m). This intermediary function comprised – among other things – the mapping of the different message schemas of the various standardization initiatives that evolved. Motivated by the success of peer-to-peer-networks in the consumer realm, some researchers explored technologies, and even business models, for peer-to-peer based networks for data exchange in the industrial domain. Technological aspects of peer-to-peer data ecosystems, such as context exchange among different world views of organizations [51] or automation of data mappings in heterogeneous settings [61], have been investigated since the late 1990s. In 2005, Franklin et al. [43] noted the growing richness of digital media and proposed that users should be enabled to create their own "data space", where a free collection of data and media objects could be managed under a user specific network of semantic metadata. In 2010, the notion of the "data lake" was coined [72] and quickly received attention in the practitioners' community. Furthermore, some researchers investigated the role of data within platform based ecosystems [59, 68, 82]. More recent research has dealt with the upcoming phenomenon of data platforms, mainly encouraged by the discourse around big data [13, 31, 59]. These studies focus mainly on platform architecture technology and data flows.

4.3.6 Research needs

While the emergence of data ecosystems offers new opportunities for the different ecosystem participants, many social, environmental and business challenges have to be addressed in order to pave the way for these opportunities to materialize. Among the most significant challenges are:

[6] Section taken from Otto and Jarke 2019 [89]

- Trust: New methods are needed to increase trust in data sharing so that more data would be available for new applications. What is needed is a framework that includes building blocks for data sharing, data management, data protection techniques, privacy-preserving data processing and distributed accountability and traceability. In addition to providing technology for platform developers, the framework should provide incentive and threat modelling tools for data sharing business developers and strategists, who consider opening data for new cooperation and business.
- Data Sovereignty: The framework should also support compatibility with the latest and emerging legislation, like the EU's General Data Protection Regulation (GDPR) and free flow of non-personal data, as well as ethical principles, like IEEE Ethically Aligned Design. This will increase trust in industrial and personal data platforms, which will enable larger data markets combining currently isolated data silos and increase the number of data providers and users in the markets. The result should aim to be platform-agnostic to be applied in multiple domains with platforms based on different technologies.
- Interoperability: The main objective should be to support a trusted data ecosystem providing easy-to-use privacy mechanisms and solutions that guarantee citizens and business entities can fully manage data sharing and privacy. The challenge is thus to provide a corresponding overall technical architecture which needs to take into account the key reference platforms and technologies to support data sharing, to improve existing solutions and architectures, to define the overall reference architecture, and to design platform-agnostic trusted data sharing building blocks and interoperability.
- Data Governance: Data Ecosystems highly depend on access to data and interactions of actors providing or using data or similar resources such as application programming interfaces (APIs). The role of data governance in these complex networks between organizations is an under-researched field. There is a lack of concepts and mechanisms to mandate responsibilities among participants of a data ecosystem. It is essential to study inter-organizational mechanisms that allow participative interactions, incentives to influence the dynamics and evolution of the ecosystem.
- Compliance with Antitrust Legislation: To avoid the risk of data monopolies, the following needs to be ensured:
    - Improving the mobility of non-personal data across borders in the single market, which is limited today in many member states by localization restrictions or legal uncertainty in the market;
    - Ensuring that the powers of competent authorities to request and receive access to data for regulatory control purposes, such as for inspection and audit, remain unaffected;
    - Making it easier for professional users of data storage or other processing services to switch service providers and to port data, while not creating an excessive burden on service providers or distorting the market.
- Data Economics: Data business, i.e. viewing data as an economic asset, will bring additional motivation for data providers and owners to open up their data for various applications. Personal data is becoming a new economic asset class, a valuable resource for the 21st century that will touch all aspects of society. The rapid development of the Personal Data Service (PDS) market will provide big changes in the way individuals, business and organizations deal with each other, as individuals assert more control over their data or service providers process personal data.

4.3.7 Data Infrastructures

Definition. *A data infrastructure is a distributed technical infrastructure consisting of components and services which support data access, storage, exchange, sharing and use according to defined rules.*[7]

The **International Data Spaces (IDS)**[8] initiative aims at data sovereignty for businesses and citizens in Europe and beyond. The IDS Association (IDSA) provides a reference architecture that enables an ecosystem for the sovereign exchange of data with clearly defined usage rights. The reference architecture defines a technical infrastructure and includes contractual regulations: at the semantic level, data linking, or analysis can technically be prevented or made possible. In this way, the classic structure of cloud services is also embedded in an interoperable digital economy with full data sovereignty on the digital infrastructures of third parties. The IDS standard solves a market obstacle: In order for data to unfold its value creation potential, it must be described and tradable according to a global and interoperable standard. This has never existed before. But DIN SPEC 27070 (to be published in Nov 19) is the first global and interoperable standard. 100 member institutions from the EU as well as from Brazil, Canada, China, India, Japan and the USA are involved; they come from all branches of industry and have developed information and governance models in the IDSA as the basis for the IDS architecture and its data sovereignty standard. More than 50 concrete application scenarios and first products are now available from companies of all sizes and sectors – together they are working on operational concepts for a sovereign data exchange infrastructure. The certification scheme "IDS_ready" also enables companies outside the association to participate in secure, IDS-based value-added processes via certified participants and components. Reference implementations and sample codes are available for developers and can be tested in testbeds. IDSA is in continuous coordination with global initiatives (Industrial Internet Consortium, OPC Foundation, Robot Revolution Initiative, BDVA) and participates in EU research projects to anchor IDS architecture and data sovereignty standards within European digitization strategies. In 8 countries, there are contractually bound IDSA hubs that bring standardization and adaptation of the technology to the respective country.

4.3.8 Questions for further research

From this brief review of literature in the field, but especially our conclusions during the seminar, we draw three conclusions:

1. The basic mechanisms of platform markets are well understood, especially the basic effect of network effects and complementing value creation in multi-sided markets.
2. Platform openness has derived as a key variable in the systematic design of a platform ecosystem. Literature suggests some factors that constitute openness that can be utilized for the design of a data ecosystem.
3. All existing analysis, however, has been conducted in the field of either consumer electronics or information industries (gaming, search engines, social media sites). Dedicated research in the context of industrial data applications is missing.

[7] For the source of this definition, please see https://www.bmwi.de/Redaktion/DE/Publikationen/Digitale-Welt/das-projekt-gaia-x.pdf?__blob=publicationFile&v=18 (in German, English version to follow very soon).

[8] For the source of this section, please see https://www.internationaldataspaces.org/wp-content/uploads/2019/10/IDSA-digital-summit-international-statements.pdf

Future research in the field is required in various aspects, which can be structured in four layers of analysis, as suggested by Gawer [49]:

- **Interfaces**: On a technical level, the openness of APIs and other technical interfaces is not just a question of programming and quality control, but first an important design factor for the ability of connected asset to perform predictive and prescriptive functionality, i.e. to enhance its capability in this regard by getting access to data also from other actors. From the perspective of a platform, the openness of an API is a signal of willingness to share data and knowledge, hence potentially attracting third parties. At the same time, open interfaces are not just a technical risk, but also reduce the ability of the originator of the data to capture unique value from this data and hence differentiate it from other market players. Research needs to investigate the choices made by companies on different levels of a platform to understand the decisions and their consequences with regard to interface design. This also includes the deployment of Software Development Toolkits (SDK) by a platform operator, which determine the ability of third party application providers.
- **Capabilities**: This layer deals with capabilities that organization need to acquire to position themselves in an data ecosystem and corresponding platforms. These capabilities include business model innovation, mastering organizational change, or capabilities of orchestrating a manufacturing ecosystem. This research builds on a rich literature of capability building and organizational sense making and will study how dedicated capabilities link to firm performance. Of particular interest are question of counterbalancing (the lack of) capabilities of one actor by capabilities of another actor on a platform. This leads to a re-interpretation of the central economic question of the boundaries of a firm.
- **Organizational design** refers to the design of a platform ecosystem and the design factors ("business model patterns") that allow for value creation and capture in these industrial data platform. A particular focus of this research will be on the level of value capture, i.e., on mechanisms that allow the different actors of a platform for profit from their participation and contributions of the platform. Building on the last work package, this also asks the question whether firms shall joining an existing ecosystem (and of, which one and under which conditions) or try to orchestrate an own?
- **Governance modes**: This final layer integrates the previous work on platform-based value creation and value capture from an external, but also an internal perspective. A central construct in this work is the degree of openness vs. desire for control of each actor in the ecosystem. Future research needs to identify possible governance modes (and patterns of platform governance) and match those to the performance of observed use cases. This should help us to understand the choices made by managers in setting up these governance modes and will (for example, game-theoretical) model the theoretical consequences of the choices under given contingencies.

4.3.9 Outlook: How to compete when all become the same (have access to the same data)

The use of networked and intelligent production systems and dynamic value creation networks in the context of a data ecosystem accelerates a faster exchange of "best practices" across corporate boundaries. For instance, this can be attempts to the process optimization in production networks, but also access to the same complementary service of a platform provider. Thus, operational efficiency can be increased for an entire industry. However, competitive advantages do not result from operational efficiency, but from strategic uniqueness! As a result, existing business models will be challenged. Companies that are highly focused on operational efficiency are facing increasing competitive pressure. Efficiency gains cannot be narrowed down to just one company. Instead of being entirely focused on operational efficiency, business models for data ecosystem must therefore initiate new differentiation opportunities for companies. The options for a more efficient and scalable custom performance have already been addressed. Openness and transparency can also become a differentiating feature: Not only during the innovation process, companies need to act more open instead of cutting themselves off. This also includes an intelligent and transparent handling of data. This intelligent handling of data enables the development of differentiating potentials using customer-specific knowledge. For example, this can be used to generate additional services or new products that satisfy the customer's benefit even better. Transparency and fairness can thereby become a competitive advantage. For companies, the challenge is to reach and maintain the competitive position as leading innovator in a specific industry. This can only happen with the best possible mix of open innovation processes and internal innovation. Companies need to find an implicit or explicit trade-off with regard to the openness of innovation processes and the type of shared knowledge. This includes finding strategies of how to implement innovations faster to the companies without internal barriers causing significant delays.

5 Use Cases

5.1 Use Cases from the Medical Domain

Sandra Geisler (Fraunhofer FIT – Sankt Augustin, DE), Maria-Esther Vidal(TIB – Hannover, DE), Elda Paja (IT Univerity of Copenhagen), Maurizio Lenzerini(Sapienza University of Rome, IT), Paolo Missier (Newcastle University, GB)

In the health care domain a variety of use cases exist, where data ecosystems are beneficial and open up new opportunities via data exchange. In the following, we describe two use cases which build complex data ecosystems in detail and analyze the challenges which are posed by them and other use cases in the health domain to data ecosystems.

5.1.1 Use Case: Multi-Site Clinical Tríal

In this example application for a data ecosystem, in the course of the research project SALUS (Selbsttonometrie und Datentransfer bei Glaukompatienten zur Verbesserung der Versorgungssituation)[9] funded by the Federal Ministry of Health, a multi-site clinical trial with glaucoma patients is conducted over one year in Germany where about 2000 patients will be included in the study starting in 2020. Glaucoma is a chronic disease of the eye with various causes possibly leading to irreversible damages of the eye nerves. For glaucoma patients it is crucial to keep the intra-ocular pressure in certain bounds to not exacerbate the condition which may lead to blindness in the worst case. The usual diagnostics require the regular creation of a status quo of the patient's condition, including an intra-ocular

[9] https://www.ukm.de/index.php?id=innovationsfondsprojektsalus

pressure profile over two successive days. For this procedure the patient has to be admitted to hospital. In the SALUS trial the advantages and disadvantages of using a mobile device at home to create a pressure profile over one week, called self-tonometry, is compared to the method applied in hospital. Additionally, the patient is equipped with a 24-hour blood pressure device, and at examinations questionnaires will be completed by the patient.

5.1.2 Study Process

Patients are included in the study by a local ophthalmologist who will explain the study to the patient and assign her to either the self-tonometry group or a control group randomly. Successively, the patient will be sent to a local hospital which will do the screening examinations and a study nurse will train the patient in using the self-tonometer. After the examination, one week of self-tonometry or a two day hospital stay, respectively, follow. The current therapy is adapted by the local ophthalmologist based on the results, if necessary. At regular intervals follow-up examinations are executed by the local ophthalmologist. After 12 months a final examination is done in the local hospital. After the trial, the collected study data will be evaluated by a research institute combining it with additional data provided by the health insurance companies of the patients.

5.1.3 Data Processing

During and after the trial all collected examination data, device data, and images of each patient have to be available especially to the local ophthalmologist, but also to the local hospital, treating this particular patient. They need the data to analyze the current condition and adapt the therapy where necessary. Hence, a glaucoma health record is maintained for each patient at a central server. Data from the self-tonometry and the 24-hour blood pressure monitoring device is transferred via a mobile device to the corresponding device manufacturer's portal. From there it is transferred anonymized to the trial server. The data and images from the examinations by the ophthalmologist and hospitals are also transferred from their systems and devices via a web application in an anonymized way to the central server. After completion of the trial, the data from the trial server has to be integrated with data of the health insurance companies and transferred to enable a comprehensive analysis by a research institute.

An overview of the use case, the involved nodes and stakeholders, and the corresponding dependencies aligned with the DE model from Section 4.2 are presented in Figure 7.

5.1.4 Challenges

We identified multiple challenges for this use case, which may also be generalized to other data ecosystems in the health domain. By the time writing, the project is still in an early stage, but where applicable, we sketch strategies for tackling the challenges.

Data views and access control. For each stakeholder, different views on the data have to be provided, as every role in this complex scenario has different access rights to the data. Hence, a very fine-granular definition of access control rights has to be provided to respond to this requirement. The data from the local hospitals, local ophthalmologists, and the device manufacturer services has to be anonymized and integrated to provide the data for the different views mentioned at different points in time during and after the trial.

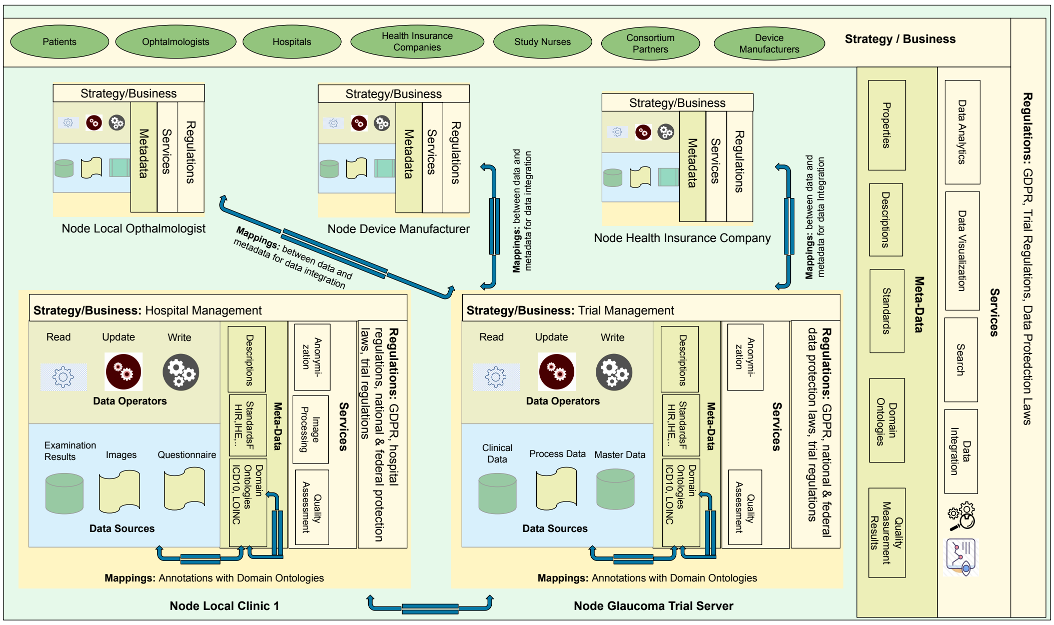

Patients
Ophtalmologists
Hospitals
Health Insurance Companies
Study Nurses
Consortium Partners
Device Manufacturers
Strategy / Business
Strategy/Business
Metadata
Services
Regulations
Node Local Opthalmologist
Node Device Manufacturer
Node Health Insurance Company
Mappings: between data and metadata for data integration
Mappings: between data and metadata for data integration
Mappings: between data and metadata for data Integration
Strategy/Business: Hospital Management
Read
Update
Write
Data Operators
Examination Results
Images
Questionnaire
Data Sources
Descriptions
StandardsF HIR,IHE,..
Domain Ontologies ICD10, LOINC
Meta-Data
Anonymi-zation
Image Processing
Quality Assessment
Services
Regulations: GDPR, hospital regulations, national & federal protection laws, trial regulations
Mappings: Annotations with Domain Ontologies
Node Local Clinic 1
Strategy/Business: Trial Management
Read
Update
Write
Data Operators
Clinical Data
Process Data
Master Data
Data Sources
Descriptions
StandardsF HIR,IHE,..
Domain Ontologies ICD10, LOINC
Meta-Data
Anonymi-zation
Quality Assessment
Services
Regulations: GDPR, national & federal data protection laws, trial regulations
Mappings: Annotations with Domain Ontologies
Node Glaucoma Trial Server
Properties
Descriptions
Standards
Domain Ontologies
Quality Measurement Results
Meta-Data
Data Analytics
Data Visualization
Search
Data Integration
Services
Regulations: GDPR, Trial Regulations, Data Protedction Laws

Figure 7 Overview of the Multi-Site Clinical Trial.

Data Integration. Thus, we need an integrated patient-centric global schema for a glaucoma health record where all data from each patient identified by a global identifier is combined. It has to be considered, if this integration should be virtual or materialized, and, how the mappings between the global schema and the sources can be created. In this use case most likely materialized integration will be used as the main data sources will not change frequently. Further integration of external data, such as medication information or usage of standards for enrichment and documentation, is a challenge given that suitable sources have to be found, their quality has to be checked, and the corresponding mappings have to be created. In this scenario various heterogeneous data sets and standards are involved which makes the integration a challenge. A data lake as basic data management architecture for the health record could be a solution to integrate various heterogeneous data sources and also allow for complex meta-data management, data quality management, access control, and search on the data. For each party/view a separate data mart could be created as a subset of the data, strictly controlling the access to the data. But such an architecture would imply a high implementation overhead as mature data lake systems are still not the usual case. A more simpler solution to be flexible according to the data storage could be also a single NoSQL database, where meta-data management and data quality management have to be realized separately.

Consent management vis-a-vis the GDPR. A further challenge is the implementation of consent management and usage control in compliance with the GDPR. Corresponding means for the deletion or editing of data at one or multiple available sites have to be implemented if the patient revokes the consent or asks for corrections. Furthermore, the GDPR also demands for transparency of operations on personal data, which requires a form of provenance tracking and auditing. Data protection and secure communication also have to be ensured, when data is exchanged between the nodes. Finally, quality monitoring during the trial can be implemented to get high value data.

5.1.5 Use Case: Precision Medicine and Health Policy Making.

In this case study, the focus is on the integration of structured and unstructured data ingested from clinical records, medical images, scientific publications, or genomic analysis. The application of a network of data ecosystems is illustrated in the context of the EU H2020 funded project iASiS [10] which aims at exploiting Big data for paving the way for accurate diagnostics and personalized treatments. iASiS[11] is a 36-month H2020-RIA project that has run from April 2017 to March 2020, with the vision of turning clinical and pharmacogenomics big data into actionable knowledge for personalized medicine and decision making. iASiS aims at integrating heterogeneous Big data sources into the iASiS knowledge graph. As input of the problem, we have a set of myriad sources of knowledge about the condition of a lung cancer patient. Electronic health records (EHRs) preserve the knowledge about the conditions of a patient that need to be considered in order to have effective diagnoses and treatment prescriptions. Albeit informative, EHRs usually preserve patient information in an unstructured way, e.g., textual notes, images, or genome sequencing. Furthermore, EHRs may include incomplete and ambiguous statements about the whole medical history of a patient. As a consequence, knowledge extraction techniques are required to mine and curate relevant information for an integral analysis of a patient, e.g., age, gender, life habits, genotypes, diagnostics, treatments, and family medical history. In addition to evaluating information in EHRs, physicians rely upon their experience or available sources of knowledge to identify potential adverse outcomes, e.g., drug interactions, side-effects or drug resistance. Diverse repositories and databases make available relevant knowledge for the complete description of a patient's condition and the potential outcome. Nevertheless, sources are autonomous and utilize diverse formats that range from unstructured scientific publications in PubMed[12] to repositories of structured data about cancer-related mutations. Various services need to be implemented in order to transform relevant data that come in different formats into a common format and for data anonymization. Additionally, since the same concept can be identified with different identifiers, the detection and representation of mappings between data sets is necessary. Each data provider establishes a regulation about the type of operators and services that can be executed over each data set; they can also indicate which of the users of the data ecosystem can execute the operators and the services. Finally, GDPR regulations need to be respected ensuring that personal data is used according to the consents provided by patients. Figure 8 illustrates an overview of a solution precision medicine use case using a network of data ecosystems.

5.1.6 Challenges

Clinical data is usually stored in diverse formats, e.g., in notes in Clinical Records, gene sequencing panels, or medical images. Additionally, these data sources may suffer from potential biases, ambiguities, and noise. To overcome these data issues, distinct knowledge extraction methods need to be included as part of the services of the network. Typical extractions methods include:

i) *Natural language processing (NLP)*;
ii) *Visual analysis and image processing*; and *Genomic Analysis.*

[10] http://project-iasis.eu/
[11] http://project-iasis.eu/
[12] https://www.ncbi.nlm.nih.gov/pubmed/

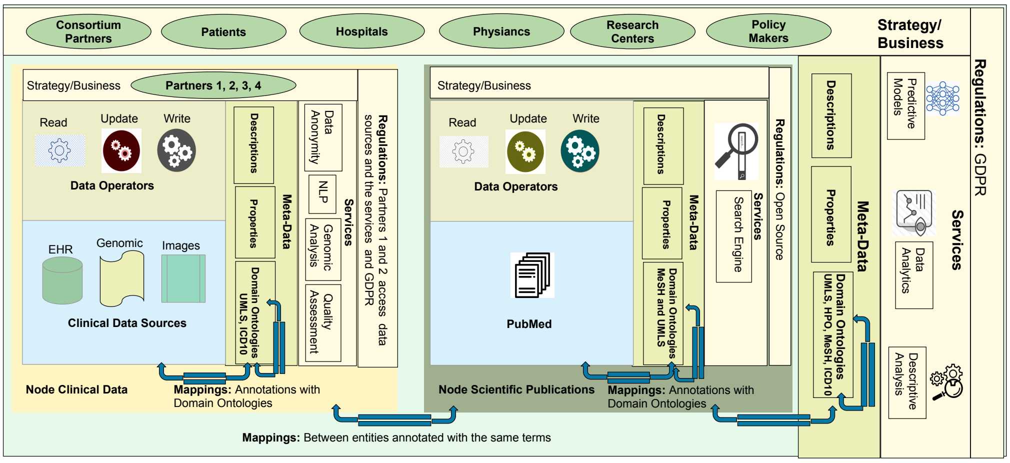


Figure 8 Overview of a Biomedical Data Ecosystem. A Network of Data Ecosystems for the Precision Medicine Use Case; it comprises two nodes of clinical data and scientific publications.

Additionally, data quality assessment techniques are required in order to identify data quality issues, as well as to define strategies for data curation. Data integration taking into account the meaning of the biomedical concepts is also challenging as the consequence of the variety of formats and representations. Moreover, exploring and visualizing data ecosystems require the implementation of scalable methods able to traverse a variety of data sources. Finally, privacy and data access control techniques are demanded to enforce regulations imposed by data providers or data protection authorities. The project aims at developing data management techniques that enable for the transformation of unstructured into a unified knowledge base. Domain ontologies like UMLS, are used to annotate the extracted concepts; these annotations provide the basis for entity matching and data integration. A unified schema is utilized to describe the integrated entities and a federated query engine supports the exploration of the knowledge base. Moreover, services for data analytics and prediction allow for the discovery of novel patterns and associations that may explain the survival time and disease progression. Initial results suggest that these knowledge discovery techniques on top of a knowledge base have the power of uncovering relevant patterns in the integrated clinical data. [71, 111].

5.1.7 General Challenges in Health Applications

5.1.8 Multitude of Stakeholders and Patient-centric Data Exchange

In the health domain many different stakeholders with diverse interests, rights, and responsibilities exist. In contrast to other domains, not only organizations are involved in use cases for DE, but also single persons, such as patients and physicians. In many cases personal data is involved which poses many challenges to the data ecosystem especially in terms of data access and data control. The role of the data owner may be taken by a party (e.g., the patient) different from the one who controls access to the data (e.g., the physician or clinic). This opens up many questions as data owner and "data controller" need to clearly negotiate which (transitive) rights the data controller may have and rules on how data exchange and usage is handled need to be defined. Though the handling of personal data is quite strictly and clearly regulated by laws, such as the GDPR, it is not an easy task to define a framework which suits all possible use cases and stakeholders' needs. A patient-centric data ecosystem would need to enable transparency of the data exchanged and of the operations executed on them, making data provenance and auditing come to the fore. A patient-centric DE should give a patient the opportunity to control what is exchanged with whom and with what constraints. Finally, it should convey trust, that data is handled correctly to the rules and constraints negotiated in the DE. Solutions to implement access and usage control for DE have already been proposed, e.g., in the International Data Space [87]. These differ in various aspects, such as the invasiveness for the internal systems (which is a crucial point for health domain use cases) or the policy languages they are using.

5.1.9 Data Anonymization vs. Data Usefulness

A further concern is the trade-off between possible anonymization of personal data and its usefulness. Depending on the use case, anonymization of the data is a must, e.g., in clinical trials and the analysis of the results, or it is desired by the providing party to not disclose certain data. But the anonymization may reduce the usefulness of the data for further analysis. Hence, it is interesting to analyze the trade-offs between the extent of anonymization and the resulting usefulness. These can be used to check against the requirements of an application intending to use the data. Usefulness and anonymity could be criteria to search for suitable data sources in a data ecosystem.

5.1.10 Data Variety, Standards, and Interoperability

In the health domain, a huge variety of data formats is used, many of which do not apply recognized standards, but rather are proprietary to specific devices or systems. Although a multitude of standards exist in the health domain, these are not used consistently and for many crucial concepts, such as Electronic Health Records, a general standard is still missing. In these cases a DE has to demand a precise documentation and meta data description of the data provided to make the data interoperable for other parties. In the first (professional health services) and second health market (consumer health and well-being services) many insulated applications and data silos exist. Applications are just written for a specific purpose or device, but are not able to connect to other systems or even export into a standard format. To make this data reusable in other contexts, other services need to encapsulate it or wrap it. Hence, a DE should provide means to easily integrate these data sources into a DE, e.g., enable services, which transform the data into an accepted standard or (semi)-automatically add semantic annotations. For these services extra costs for data consumers or providers could be charged.

5.1.11 Meta-data Usage, Quality, and Mappings

Many medical taxonomies and ontologies exist to express meta-data and to annotate data. This is beneficial on the one hand to make the data more “understandable” and enables more intelligent applications, which also consider the context and meaning of the data. On the other hand, many of the most used vocabularies are so huge, such as MeSH, SNOMED, or the NHI Thesaurus, that working with them and maintaining them is very difficult. This results in quality issues in the vocabularies and, as a consequence, also in the applications that are based on them or the mappings created between two or more ontologies. Also data curation as such is an error-prone and tedious process, leading to poor annotations when this is performed by untrained persons or by algorithms.

5.1.12 Data Quality

Data quality (DQ) plays a crucial role in all data-intensive applications, hence, also in medical applications. There exist several sources of DQ problems in health applications. Much of the health data is recorded or transformed using electronic devices and sensors which may have an inherent physical imprecision or are easily prone to failure and lead to DQ problems. The monitoring of DQ for online or streaming sources is difficult and may lead to additional problems, such as synchronization and timeliness problems. Additionally, a lot of data is collected manually, e.g., using paper forms, which are also prone to errors, missing data, or transfer errors when digitizing the data. Furthermore, a multitude of data sources and formats may be involved in just one application. The conversion between formats, the integration with other data into existing data sets or databases, and the aggregation of data may as well lead to DQ problems. Many standards, for example FHIR, provide guidance to structure the data accordingly, but it can be very different how the guidance is implemented which may also lead to problems in the data format and interpretation. Also, the semantic consistency of data (is the data representing the patient and her condition correctly?) of single and between multiple examinations and measurements must be checked which often has to be a manual task. It is interesting to investigate how integration of several sources and Machine Learning algorithms could be used to support this process.

5.1.13 Data Protection and Data Sovereignty

Working with personal data is highly regulated by laws on different levels. For example, considering Europe and the EU, data security and protection is regulated by the General Data Protection Law (GDPR). Each country in the EU additionally may have a national law (e.g., in Germany the Federal Data Protection Law), which extends the directly applicable GDPR. Further regulations can be defined on state level (e.g., the Krankenhausgesetz in Bavaria or the Data protection Law in North-Rhine Westphalia). These multiple overlapping regulations lead to a highly complex environment for data ecosystems and data exchange. In turn, this requires a highly flexible framework which must combine usage control, access control, data provenance, and further means to enforce data protection and enable data sovereignty.

Data sovereignty for patients is desirable, but not easy to implement. Patients may not feel capable of deciding which information should be accessible to whom. It is hard to decide for laymen, what exactly is sensitive data and if this data is crucial to a specific health professional to enable the best possible treatment. The grade of transparency regarding data exchange may also be subject to discussion, as for example a very “chatty” notification protocol may annoy or make patients feel not secure.

5.1.14 Conclusion

In this section we have presented two use cases which constitute complex DE in different ways, but sharing similar problems. In general, the challenges for data exchange are manifold and specific in the health domain as many stakeholders with many different interests are involved. Especially, single persons, such as patients and doctors, play a major role in the DEs which requires to overthink mechanisms so far identified and implemented for industry and companies. Highly sensitive data and the corresponding ethics and laws around it, pose crucial challenges to DEs. And finally, the high variety in data formats and metadata definitions make the domain description and application implementation special. Hence, we postulate that a framework for DEs has to take all of these challenges and requirements for the health domain into account to propose general solutions.

5.2 Industrie 4.0 Data Ecosystems Examples

Egbert-Jan Sol (TNO – Eindhoven, NL)

Multi-sided markets are based upon an alliance agreement for sovereign data exchange between organizations and result in a distributed data driven ecosystem. In such data ecosystems (DES) each organization can do more then when using only its own data, and once properly architectured this can be realized without the need to see all data such that each party remains control over its own business.

Industrial data ecosystems are encountered in the (discrete) manufacturing, the (chemical) process industry, in (maintenance) and logistics services, but also in amongst other industries as telecommunication ecosystems as GSM. In practice all these industrial cases there is a physical device that can get a so-called digital twin, being the digital representation of a physical good. With this digital twin more data can be linked to the device resulting in data sets that can be (partly) shared. In industry we encounter data sets that consist of sensor data, copyrighted data as drawings all the way up to very valuable, confidential data sets.

We introduce in this section four notions: 1-a data analogy of material, 2-digital twins and identifiers, 3-four classes of data and finally 4- a function of controlled data exchange. It is followed by four industrial use cases of data ecosystem: GSM mobile telephony, discrete manufacturing, traceability of food/goods, and preventive maintenance with hints to aspects as use of an association, international standard and cyber security protection consequences.

5.2.1 From data element to data sets with sovereignty control to value

Data records, due to their digital nature, are often seen as discrete elements. But to accept that certain data is more valuable than other data, an analogy with the material/process industry might be more applicable.

Say, an individual sensor reading could be seen as atom and a long set of sensor readings as base material. Sensor data as such is not so valuable and sensor data is not copyright protected. Then if one combines iron with carbon to produce steel or mix a polymer with a color ingredient one can get a stronger alloy due to the iron/carbon matrix/lattice or a colored plastic. Combine two data set also leads to a more valuable set then just a single list of data points, just as steel is more valuable than its base materials. Next image the combination of a drawing and a set of manufactured parts that are verified to be produced from the steel alloy or plastic within specs according to the drawing. This set is more valuable than just a bag of polymer granules or a block of steel. Data sets that are verified are also more valuable.

We won't identify the single sensor data element as we won't identify an individual atom. The base material is already more valuable, but not as valuable as a strong alloy. Similarly we might identify a list of sensor points having some value, but the combination with a safe boundary set within the sensor points must stay is already more valuable. And an end product similar as a list of customer bills is even more valuable.

There is an analogy in data set similar to atoms, to base materials to chemicals, to discrete products up till a product owned by someone. Single data lists, data sets and their relations and finally a whole data bases can be have a certain value for a user. And combining data sets over multiple parties and share it in a larger data ecosystem can even be more valuable. The challenge is to construct the data ecosystem such that parties can do more with the

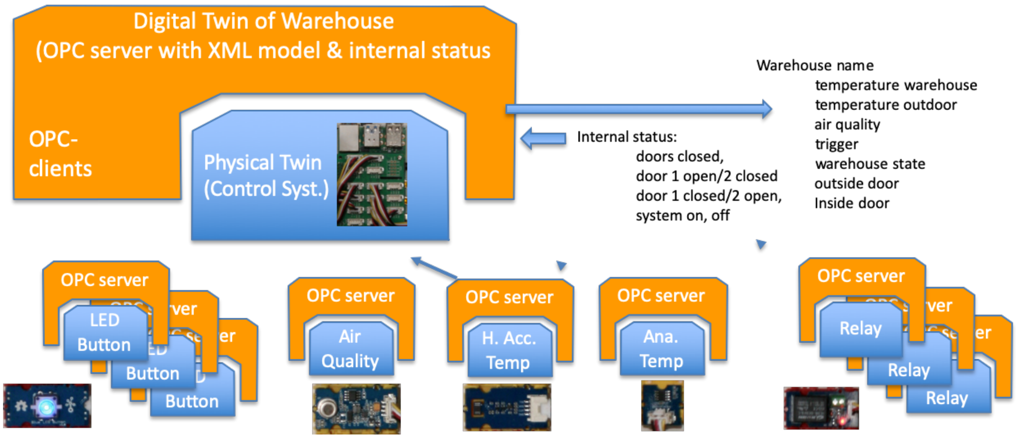


Figure 9 Nested Digital Twin with data elements.

available data without the need to see all data. Similar as you driving a car without knowing and having the access to all the internals, the manufacturing secrets as costs of components.

But whereas a product has an owner, a manufacturer, a user, etc., data can be copied over and over again. Here data ownership is more difficult to define. Only copyrighted data, i.e. where human creativity is involved as e.g. drawings, can have the copyright owned by a legal entity. And for certain data sets there exists a databank act. For private data Europe has the notion of data sovereignty in the form of the GDPR. As of today, in industry there is no legal binding concept a data ownership and/or sovereignty. There only exist contractual relations on the sharing and use of data between parties. Only the producer of the data has sovereignty control over whether to share the data with others. But e.g. sensor data has no copyrights and once shared without a contract other can do what ever they want.

5.2.2 Industrial data twins and data identifiers

Each physical object can have a digital representation in the form of a data representation as simple as just a number or as complex as a large data set with drawing, manufacturing recipes and use history. In general, it will have a name and a digital identifier too. In figure 9 a physical object or asset with its digital twin as a kind of administrative shell around it is symbolized.

An object can consists of subpart or a group of objects can form a larger object. This results in a hierarchy of digital twins each with their own data and links to other digital twins. In particular similar objects could each share the same design and their own instantiation. In that case, these digital twin instantiations (DTI's) share a common design (see figure 10 with the DT on the left and the DTI's on the right).

The design part of a digital twin could be a set of drawings, software, recipes etc. This design and the digital twin (of the design) (DT) is owned by a legal entity. And the physical objects can be owned by (other) legal entities. But what about ownership of the digital twin instantiation of each object with e.g. the historical (sensor and use) data?

5.2.3 Confidential, shareable, shared and public data sets

Each digital twin of an object, a system or even an organization can have confidential information that is not (to be) shared by anyone else. The owner of an object can also decide that certain data can be shared with other legal entities with whom the owner decides to

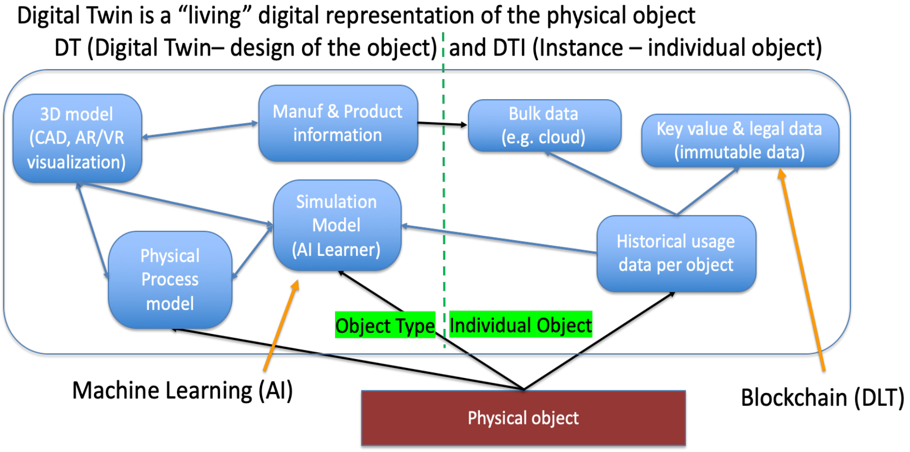

Digital Twin is a "living" digital representation of the physical object
DT (Digital Twin– design of the object)
and DTI (Instance – individual object)
3D model (CAD, AR/VR visualization)
Manuf & Product information
Bulk data (e.g. cloud)
Key value & legal data (immutable data)
Simulation Model (AI Learner)
Physical Process model
Historical usage data per object
Object Type
Individual Object
Machine Learning (AI)
Physical object
Blockchain (DLT)

Figure 10 Digital Twin with its DT and DTI (instance).

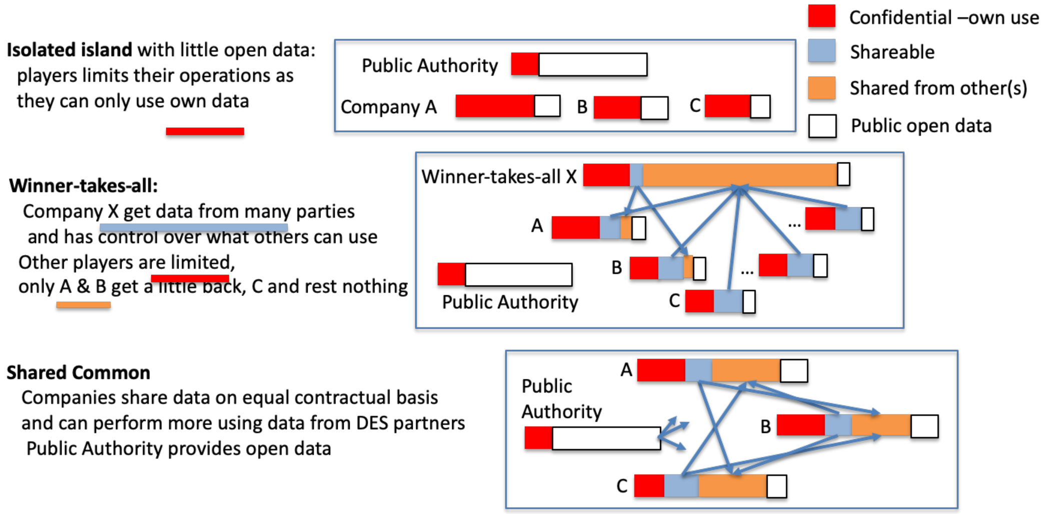

Confidential –own use
Shareable
Shared from other(s)
Public open data
Isolated island with little open data:
players limits their operations as
they can only use own data
Public Authority
Company A
B
C
Winner-takes-all:
Company X get data from many parties
and has control over what others can use
Other players are limited,
only A & B get a little back, C and rest nothing
Winner-takes-all X
A
B
C
...
Public Authority
Shared Common
Companies share data on equal contractual basis
and can perform more using data from DES partners
Public Authority provides open data
Public Authority
A
B
C

Figure 11 Anatomy of a Data EcoSystem.

share data with. This requires both a technological standard to exchange the data as well as a legal contract to specify the conditions of sharing. One can also receive data made shareable by one or more other parties, in a similar way the owners made some data shareable to others, again using tech standards and legal contracts). Finally, there might be public data that can be used too, see also figure 104 on the anatomy of three different kinds of data ecosystems illustrated as isolated islands, winner-takes-all and shared common. The last one is the preferred one in case one wants to keep a certain data sovereignty.

The result is that around every physical object, systems and even organizations a digital twin can exist with data and that these digital twins can form a data ecosystem. In industrial cases that data ecosystem will be related to physical objects, digital identifiers, data sets and relations between data sets where the digital twins can exist and be copied unlimited times in all kind of subsets while there is only one instant of the object.

Different players in a data ecosystem can have different views of the data and therefore different digital twins of the same object. Of course, by expanding the own view of the digital twin data with data shared by others a better and often more valuable digital twin can be realized.

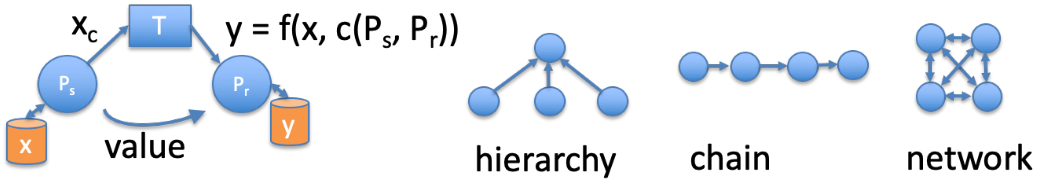


Figure 12 Different Data Ecosystem architectures.

To function in a data ecosystem, it is therefore sensible to make some or all of the own data on that DT object and/or DT instances shareable to others too. Depending on the business parties can agree to keep as much confidential (not shared) and minimize shareable data. In an extreme case one party could try to collect as much data from the others and maneuver itself in a data monopoly position. It depends on the parties how much they are willing to share, the required contracts and the technical interfaces they select and use for data exchange.

5.2.4 A Data Ecosystem function

To be able to share shareable data between parties in an ecosystem one needs, next to technical data-communication standards and legal contracts a software function. With X the data and C the legal constraint put on the data exchange, we call the entity that controls the exchange of data T (transmit). And with Ps (or S) the sending party and Pr (or R) the receiving party, T performs a function f(X, C(S, R)). T receives data X from an Ps under specified constraint instruction. See figure 12.

T can be distributed as in e.g. IDS over all parties or T can be a clearing house with processes the data as specified. Any party P could have a sending and a receiving side. In a data ecosystem there can be many S's and R's. An unbalanced situation is when there are many S and only one R, in particular if T is completely under control of the legal entity R. A tightly coupled chain might be when a flow existing from S to R follows by R to O, O to P, etc. In an hierarchical system many parties send and receive data up- and downwards and not so much side wards. Finally in an fair data-ecosystem all parties can exchange data with everybody else.

5.2.5 The GSM model

The GSM association runs a data ecosystem since the 1990-ties. With mobile telephony 1 users could not roam. With GSM a user can roam to a geographical different network of another GSM operator as in cross border travel. Because of the roaming agreement a GSM operator can provide more services then in generation 1 mobile telephony networks. But GSM operators need to share data on the roaming users. This is done by the concept of home and visitor location centers. In essence this is data-ecosystem that provides its participants more services then without it. See figure 106

In the home location center the user, the GSM nr (and device ID (IMEI)), the used service and the billing information is known. In the visitor location center only the device ID with the usage of mobile voice time and data traffic is known. This information is shared with the home GSM who then pays a bulk fee to the visited GSM operator and adds up the received usage data to the customer usage to finally produce a bill. Next to the realtime

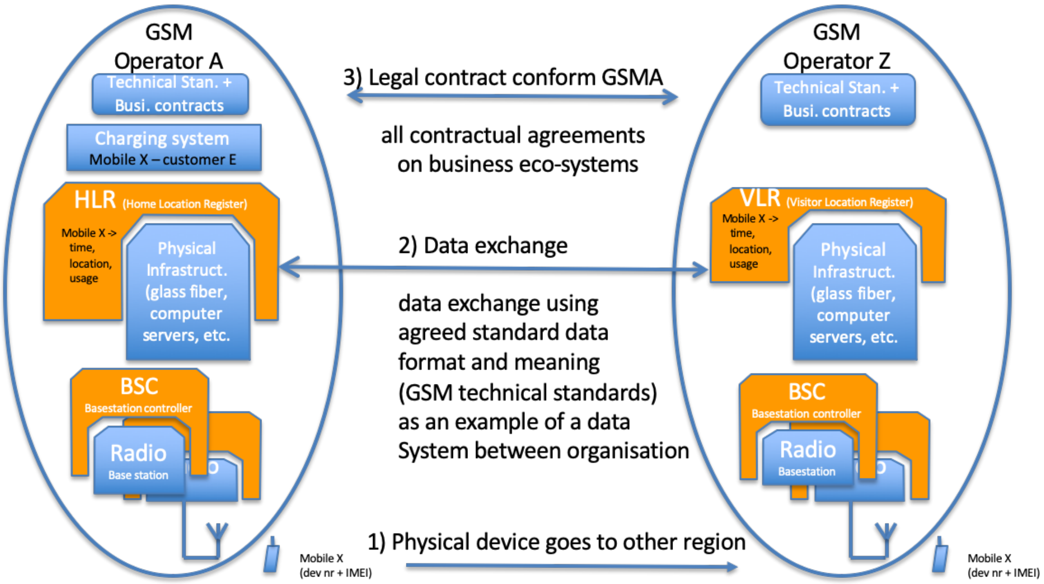

GSM
Operator A
Technical Stan. +
Busi. contracts
Charging system
Mobile X – customer E
HLR (Home Location Register)
Mobile X -> time, location, usage
Physical Infrastruct. (glass fiber, computer servers, etc.
BSC
Basestation controller
Radio
Base station
Mobile X (dev nr + IMEI)
3) Legal contract conform GSMA
all contractual agreements on business eco-systems
2) Data exchange
data exchange using agreed standard data format and meaning (GSM technical standards) as an example of a data System between organisation
1) Physical device goes to other region
GSM
Operator Z
Technical Stan. +
Busi. contracts
VLR (Visitor Location Register)
Mobile X -> time, location, usage
Physical Infrastruct. (glass fiber, computer servers, etc.
BSC
Basestation controller
Radio
Basestation
Mobile X (dev nr + IMEI)

Figure 13 GSM network.

exchange of visiting or roaming data of the device there is the contractually data exchange on the total usage of visiting devices. Notice that te home GSM operators has personal data as customer name, billing info and GSM nr. The visiting GSM operator only has the device ID. When ever a device roams it device ID indicate the home GSM operator and visiting operator informs the home location register in which visitor location the device currently is. When the mobile phone number is called (in the home location visitor), the GSM home operator routes the call to the visiting GSM operator.

In the GSM case the identifiers are the home/visitor location centres, the GSM device ID each linked data sets as customers, device usage, etc. The transport T has three simple functions: updating the device visitor location (in realtime) to the home, forwarding calls to the device in the visiting network and sending the usage at aggregated level to the me too.

5.2.6 Discrete manufacturing supply chain

This example is a first-tier supplier network where orders from the OEM-er are sent to a supplier including o.a. drawing information and where finished subparts are sent to the OEM to built the product. Without the order and, if it is not a standard component, the drawing information, the supplier cannot produce. Often parties try to keep as much data as possible confidential, but more and more data is sent to the OEM-er, often on demand of the OEM-er for traceability, quality monitoring and control of the (realtime) logistics in the chain. Sometimes an independent player T could be used to assemble business information as market share and relative quality information enabling suppliers to upgrade their performance. In other cases the OEM can give feed back information to the supplier performance.

As with the GSM example, here to digital identifiers, but also digital twins and a three level physical object, digital data exchange and legal contracts are encountered.

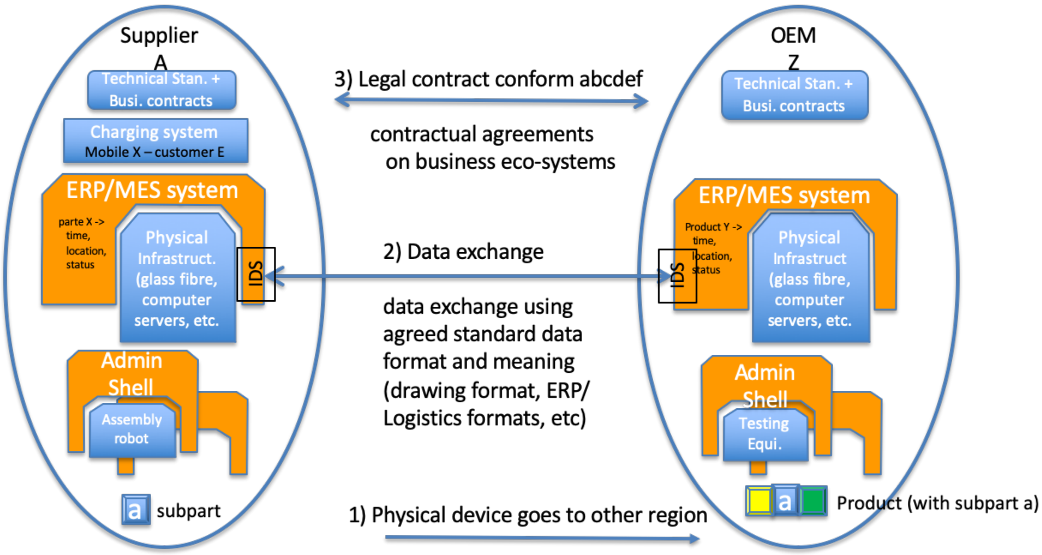


Figure 14 Supplier - Original Equipment Manufacturer network.

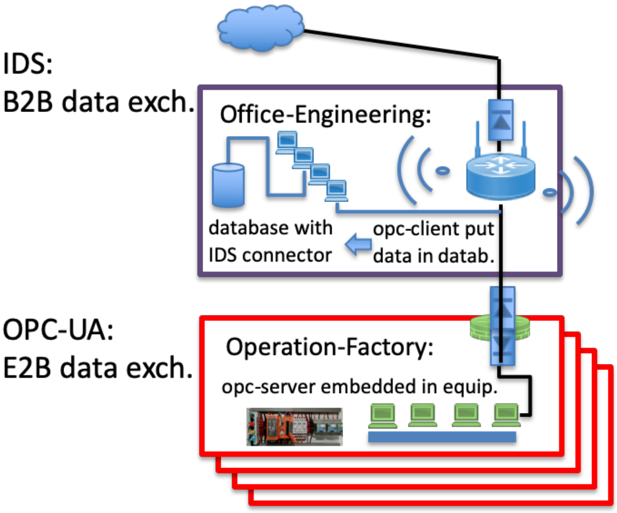
5.2.7 Food traceability

More and more industrial chains demand extensive traceability. Examples are the automotive and aircraft industries, but also the food industry. In figure 1 a warehouse with sensors was shown. In that case it is a part of chain where food is stored under certain temperature and humidity conditions. The warehouse operator uses the data to optimize the energy consumption, but the same data can also be shared or "sold" to the food owner to monitor the storage conditions for traceability and quality control. But sharing the data, with or without payments, the whole (data)eco system can perform better.

5.2.8 Predictive Maintenance

Predictive Maintenance is similar to the food traceability use case. In this case a maintenance service party monitors the condition of equiment to predict a stop for maintenance. Often the operator of the system uses the data for (real-time) control at the same time. By sharing the data with another party that party can combine more data for others to improve its prediction algorithms. The specifics of this use case is that these environment tend to require strict security and access control.

Figure 15 shows the case where OT (operational technology) has shielded data access in its own subnet and data cannot be sent directly to others. In this case the data is sent by OPC-UA (IEC standard) protocol through a firewall to a server to be stored there and made suitable to be further processes outside the operational production environment. These two level data ecosystems often restrict other legal parties to have an active proces directly communicating through the customers network to their own remote services support networks. OPC-UA is used for equipment communication within an organization. Other standards as the IDS (international data spaces) can be used for data exchange between different organization using the IDS technical and contractual agreements.



■ **Figure 15** data ecosystem inside and outside a factory.

5.3 Use Cases from the Smart Cities Domain

Cinzia Cappiello (Polytechnic University of Milan), Bernadette Farias Lóscio (Federal University of Pernambuco), Avigdor Gal (Technion – Haifa, IL), Fritz Henglein (Univ. of Copenhagen, DK & Deon Digital – Zürich, CH)

The Smart city concept has various definitions that are associated with viewpoints that range from the people perspective to a more technological perspective. Such latter point of view suggests that smart cities mainly focus on adopting the next-generation information technology to "all walks of life, embedding sensors and equipment to hospitals, power grids, railways, bridges, tunnels, roads, buildings, water systems, dams, oil and gas pipelines and other objects in every corner of the world, and forming the "Internet of Things" via the Internet"[103]. Smart Cities data collected by heterogeneous devices and data sources are strategic to perform analysis, to extract relevant patterns, to detect inefficiencies, and to propose innovative solutions to improve the quality of life of citizens.

5.3.1 Scenario description

Goal: Definition of an ecosystem that includes several cities (learning from one city can help to accelerate the process in another city).

The Smart Cities DE, presented in Figure 16, can be seen as a composition of Data Ecosystems. Each city has its own DE and they can exchange data and knowledge. Each DE is composed by Data Providers and Services. The Data Ecosystem of City 1, for example, is composed by five Data Providers, which can also play the role of a Data Consumer. Data from different providers can be used by a Data Consumer to offer more advanced service. In this scenario of data mobility, Data Provider 1 and Data Provider 3 provide bus data to Data Provider 2, which will perform time predictions based on these data. In a similar way, Data Provider 4 receives bus data from Data Provider 1 as well as time predictions from Data Provider 2. Data Provider 4 uses these data to develop a better location planning

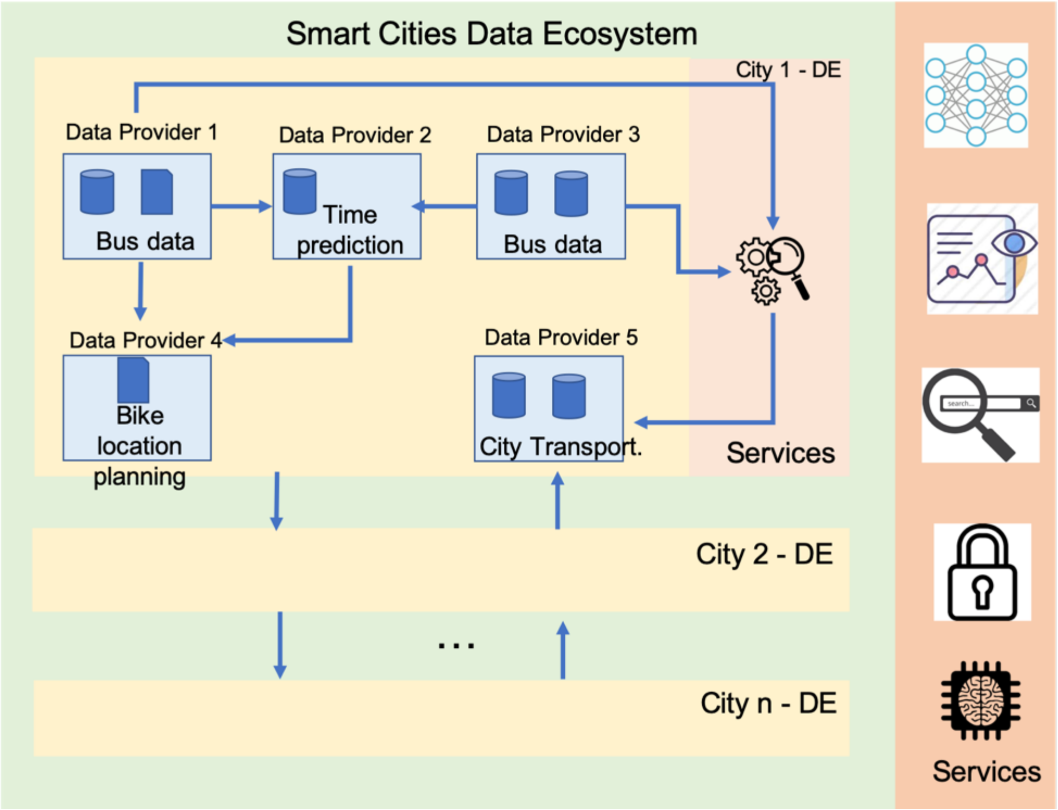


Figure 16 Smart cities data ecosystems.

for bikes to rent. Finally, Data Provider 5 uses the services offered by the DE to provide information to citizens.

In such ecosystem, we have to clarify that a city can learn from another city but it is difficult that a city might inherit or use the same applications/services of another city. In fact, most of the smart city applications are context- aware and tailored for the city needs.

5.3.2 Challenges

- General challenges
  - It is important to guarantee that people use data in the correct way: data usage constraints have to be defined.
  - Looking at the data infrastructure, it is difficult to understand which are the boundaries of a node. It might be adopted a data provider-oriented approach in which a node is a data provider within its own data sources.
- Challenges related to derived data issues
  - In the data ecosystem of a city, in general a data provider can be the owner of raw data sources or can provide derived data created transforming data gathered by other data providers. In summary we can distinguish between raw data and derived data. It is important to find a way to describe and manage derived data.
  - How do we define the usage constraints of the derived data? It might happen that raw data are regulated by usage constraints but derived data do not contain personal data and vice-versa. Besides, integration between two sources could reveal personal data.

  - The code related to the applications/services should be represented. Also the code should be protected
  - View maintenance: model improvement or data improvement.

- Challenges related to Quality assessment
  - In this architecture, exchanged data should be high quality data in order to exploit the potential data value. It is necessary to define who should be in charge of the quality assessment. In fact, data quality assessment could be centralized and in charge of the platform through a platform service also using crowdsourcing.

6 Manifesto

The Dagstuhl Manifesto is an agreed result of the Dagstuhl Seminar "Data Ecosystems: Sovereign Data Exchange among Organizations" that took place on 23-27 September 2019 in the Schloss Dagstuhl Leibnitz Zentrum fur Informatik, that joined forces of the researchers, practitioners and experts from Europe and worldwide, with the goal to discuss new challenges in building data ecosystems supporting sovereign data exchange among organisations involved into whole data value chain. We understand that:

1. Data exchange among organizations is a key enabler for the digital economy of the future.
2. A secure, reliable and performant data exchange infrastructure is a basis for operating sustainable data ecosystems supporting the whole data value network involving data providers, data consumers, and service providers.
3. Enabling assessment and awareness of data quality are core requirements in a data ecosystem.
4. Organizations and individuals must be regarded as data sovereigns entering into data ecosystems according to agreed contracts.
5. In a data ecosystem, data should be considered both an economic asset and a tradable commodity according to specified conditions of use.
6. Defining metadata is necessary for enabling a variety of operations on data and with data, along the whole data value network.
7. Defining semantic interrelationships among data sets is a key problem/task in creating scalable data ecosystems supporting effective data exchange.
8. Defining a generic data ecosystem architecture is necessary for supporting interoperability at multiple ecosystem levels, including data sources, semantics, applications, workflows and processes, governance and economics.
9. Economically viable data ecosystems architectures should leverage successful experiences from similar distributed systems such as GSM employing context roaming between independent domains.
10. Future data ecosystems should serve the public interests, in particular by supporting data related projects to address the UNDP 17 Sustainable Development Goals (SDG) .

7 Statements of the participants

7.1 Cinzia Cappiello (Polytechnic University of Milan, IT)

My research focuses on Data and Information Quality. In particular, in the last years, I addressed issues related to the Data and Information Quality assessment and improvement in the context of service-based applications, Web applications, Big Data and IoT. Data Quality is defined as fitness for (intended) use, that can be seen as the capability of a data set to satisfy users' requirements [6]. This definition suggests that quality is subjective: a data set that is appropriate for an application/user might not be suitable for another one. For this reason, currently, I am working on the design of an application-aware data quality assessment platform: data quality should be evaluated by considering the actual usage of the considered data set. Moreover, in this field, most of the literature contributions propose approaches for structured data sets. In data ecosystem, the increasing volume and variety of the data sources needs the definition of new data quality methods. Such methods should be designed by considering the type of source (e.g., data streams vs traditional db) and the type of data (e.g., text vs numerical). In fact, these two variables impact on the dimensions to evaluate and on the metrics to use. In summary, in order to enable an automatic evaluation of the quality level of the data sources it is necessary to define an adaptive data quality assessment service able to trigger the appropriate mechanisms by considering the characteristics of the sources and the application/user that requested data.

7.2 Ugo de'Liguoro (University of Turin, IT)

My interests are to logic and computation. Central are both pure and typed λ-calculus. I have also investigated other formalisms, like the ζ-calculus and the π-calculus, considered as theoretical calculi modeling object-oriented and concurrent programming languages respectively. Key methodologies in my work are type systems and denotational semantics, that nicely correspond each other especially in the case of intersection type assignment systems. The latter express functional properties of λ-terms seen as programs, and I have worked to include in the theory also non functional aspects, like non determinism and control operators. I believe that intersection types have a great potential to model both behaviour and data; an experience in that direction is the participation to a project of program synthesis from components in Dortmund, based on intersection types and combinatory logic. In a related field I have been working on session types and contract theory. This is about modeling protocol compliance of multiple principals interacting through a network, enjoying safety and liveness properties. A particular concern in this work is checking for consistency of local protocols to form a well behaved system, and the adaptation of each to the others when considered as components independently built and specified.

7.3 Yuri Demchenko (University of Amsterdam, NL)

The fact that data has a value is commonly recognised. However, data value is different from those associated with the consumable goods. There are a number of initiatives to create data markets and data exchange services. Well established business models of paid or commercial data(sets) services such as data archives are based on the service subscription fee. Quality of datasets in many cases is often assessed by independent certification body or based on peer review by expert. However this model does not provide a basis for making data an economic goods and enable data commoditisation. Another important development in making the best use of research data is based on wide implementation of the FAIR (Findable – Accessible – Interoperable – Reusable) data principles, which are widely supported by research and industry. However, emerging data driven technologies and economy facilitate interest to making data a new economic value (data commoditisation) and consequently identification of the data properties as economic goods. The following properties proposed in [32] leverage the FAIR principles and are defined as STREAM for industrial and commoditised data: [S] Sovereign - [T] Trusted - [R] Reusable - [E] Exchangeable - [A] Actionable - [M] Measurable Other properties to be considered and necessary for defining workable business and operational models: nonrival nature of data, data ownership, data quality, measurable use of data, privacy, integrity, and provenance. Defintion of the data properties as economic goods must be supported by creating consistent and workable models for data exchange and commoditisation, to facilitate creation of new value added data driven services. Consistent data pricing and data markets models are equally important for government funded and sponsored research, open data and governmental data. The proposed Open Data Market (ODM) model is based on the adoption of the IDSA Architecture and data sovereignty principle [87]. The ODM must be based on relevant industry standards and provide secure and trusted data exchange between data market actors: data producers/owners and data consum-ers, services and applications developers and operators. A functional data market model and architecture should include multiple components such as the secure trusted data market infrastructure as well as regulatory basis. The proposed open data market model would be decentralized and allow creating virtual private market instances to support data exchange between peers or group of peers. This gives the advantage that network nodes, data sellers and data buyers, receive the full benefits of the data market while retaining full control of their data. This is an important requirement for data markets for industrial data where companies, the data owners, want to retain control over their data, maintaining data sovereignty. The ODM research evaluates the blockchain technology to enable an open controllable trusted data market environment for secure and trusted data exchange, and support data value chain (provenance) and create a bias for data monetisation [13].

7.4 Elena Demidova (Leibniz Universität Hannover, DE)

Recently, we have been involved in several projects focused on analysing data for urban mobility use cases. Examples include prediction of impact of planned special events on traffic [106], identification of structural dependencies in urban road networks [107], or analysis of driver behaviour. Realisation of such use cases requires integrated analytics of data originating from different domains (map data, traffic information, car trajectories, car sensor data, city infrastructure data, population statistics, people movement, etc.), as well as different sources and owners. Currently, most of the data relevant for such analytics is locked by different organisations (e.g. companies or city municipalities), making it hard to obtain the overall picture and build accurate prediction models. In this seminar, I would like to share

[13] https://datapace.io/datapace_whitepaper.pdf

experiences we have collected in several related projects (including Data4UrbanMobility[14] [108], Simple-ML[15][53] and CampaNeo[16], leading to challenges related to collecting such data, and making use of the data in the analytical use cases more transparent to data owners.

7.5 Boris Düdder (University of Copenhagen, DK)

Boris Düdder's core interests are in formal methods of software engineering and distributed systems, for example, by employing logical methods, rewriting systems, and model checking. His interests concerning the Dagstuhl Seminar are methods for automatic configuration of exchange of sovereign production data of manufacturers and suppliers securely while guaranteeing ownership, confidentiality, and privacy. He can contribute his expertise and industrial experience in formal methods for synthesis, analysis, and verification of software for industrial production systems and supply chain (including logistics) as well as distributed ledger technology, e.g., for proof of provenance applications and finance. Industry experience includes logistics, whiteware and aircraft manufacturers.

7.6 Bernadette Farias Lóscio (Federal University of Pernambuco, BR)

In the last years, I have been involved in several projects related to Open Data and Data on the Web. One of my main interests concerns how to share data on the Web in a proper and sustainable way. The growing interest in sharing data on the Web gives raise to several challenges related to important subjects including data provenance, data privacy and data access. In order to help data providers and data consumers to face those challenges, the W3C Data on the Web Best Practices Working Group proposed a set of 35 Best Practices, which cover different aspects related to data publishing and consumption, like data formats, data access, data identifiers and metadata. I was part of the DWBP working group and I am one of the editors of the Data on the Web Best Practices recommendation. This experience was very rich and I could have the opportunity to identify several open research challenges related to Data on the Web. One of these challenges concerns the creation of sustainable Data Ecosystems.

While Data Ecosystems are gaining importance, several ecosystems are not sustainable and consequently the effort spent by their actors end up not being properly used or forgotten. The lack of communication and cooperation between data producers and consumers is one of the main obstacles moving towards sustainable Data Ecosystems. Moreover, designing, developing and further maintaining systems for Data Ecosystems are not trivial. Recently, I have advised several students in subjects related to Data Ecosystems, including a metamodel to represent data ecosystems, a metadata curation framework for data ecosystems and a framework to assess data ecosystems health. In this seminar, I would be glad to share my experiences on these subjects and I am willing to learn from the experiences of other participants.

[14] http://data4urbanmobility.l3s.uni-hannover.de
[15] https://simple-ml.de/
[16] https://www.l3s.de/de/projects/campaneo

7.7 Avigdor Gal (Technion – Haifa, IL)

My research focuses on effective methods of integrating data from multiple and diverse sources, in the presence of uncertainty. My current work zeroes in on schema matching – the task of providing communication between databases, entity resolution – the task of identifying data elements that relate to the same real-work entity, and process matching – the task of aligning process activities. In the context of this seminar, I have contributed to the design of an intelligent data lake, one in which integration is seamlessly generated and tested with the assistance of experts, heuristics and machine learning algorithms

7.8 Sandra Geisler (Fraunhofer FIT – Sankt Augustin, DE)

My field of expertise and research comprises several aspects of data management. Especially, my work is concerned with techniques in the area of data stream management, big data in general, data quality, data lakes, meta-data management, data integration, and semantic web technologies. Furthermore, I have a strong background in the medical domain and medical informatics.

Currently, I am working on projects in the context of the Fraunhofer Medical Data Space [17]. The Medical Data Space can be viewed as a distributed data ecosystem which has specific requirements stemming from the challenges of working with health data. It is especially concerned with data sovereignty, data security, and data privacy striving for compliance to data protection laws on various levels. The data in the Medical Data Space remains in custody of the data owner and is only exchanged in a controlled and secure manner. For example, data important for the follow-up care of patients after a surgery may be shared between several parties involved in the care process, such as the general practitioner, the outpatient nursing service, the hospital, and of course the patient herself. But, every party may only need a specific part of the data, or should not be allowed to access all of the data, or should only see results produced by algorithms using the data as input. Furthermore, data quality plays a crucial role as the data may be used to support processes which may have indirect or direct impact on the health of a person. Also data sources with potentially erroneous data, such as health monitoring devices, are often involved. But, controlled and automated data collection of the users may also help to elevate DQ which in consequence will also lead to a higher data value and quantity for health care services or other data consumers, such as research studies. Finally, challenges regarding the integration and interpretation of the distributed and very heterogeneous data have to be tackled to provide a use case specific and user-friendly view on the data.

Based on my research interests and current project work, I can contribute to the seminar in terms of the discussion of applications from the health care sector and their specific challenges and requirements focusing on data sovereignty, data quality, and data integration.

[17] https://www.medical-data-space.fraunhofer.de

7.9 Benjamin Heitmann (Fraunhofer FIT – Aachen, DE & RWTH Aachen, DE)

The data market places of the future will be enabled by technologies which protect the value of data for all participants of a market place. In my experience, too often the real value of data or other digital assets for a stakeholder is neglected, as processes and algorithms can currently be better protected. In order to protect the value of data in a market place, the three goals of security, privacy and data sovereignty need to be fulfilled.

Security means protecting the digital assets which an organisation controls and which might be valuable to an attacker. Examples of assets which usually are secured, include domain-specific knowledge collections such as life-science data sets, financial records, records of human resource data, product sales data, and instance data for business processes. Failure to protect the digital assets of an organisation results in losing control of knowledge, processes and customers, as competing organisations could gain access to the same assets.

Privacy means protecting the data of individual users, which is required as input for many digital value chains. Examples of private assets include digital health records, purchase histories for e-commerce sites, historical data about consumption habits for news, movies and music. Failure to protect the personal data of individual users will result in unwillingness or reduced incentives for users to share or sell their data. This will in turn disable any processes which require input data from users to generate more value for an organisation.

Data sovereignty means protecting the digital assets which an organisation or individual controls and which have been sold or shared willingly with another entity. Examples of assets which might be shared or sold without losing control include R&D data sets for manufacturing, R&D data sets for discovering new gene interactions, personal data about historic fitness activities or health indicators such as blood pressure. Failure to protect assets after sharing or selling them will result in loss of control over the purpose of using the data, and in decreased incentives to share or sell the data with additional entities in the future.

In future data market places, the data protection needs to fulfil the requirements of all stake holders at the same time: data generator, seller and buyer. In addition, usually all three goals of security, privacy and sovereignty need to be reached at the same time. This requires technologies which enable **enforceable guarantees for protecting data during processing**. In my experience, some of the technologies with the highest potential come from new advances in information processing and encryption. The most notable new developments are in the areas of secure multi-party computation (SMPC), homomorphic encryption (HE), and differential privacy, anonymisation and synthetic data generation.

Together with new architectures, business models and legal frameworks, the listed technologies for enforceable guarantees to protect data during processing can enable sustainable data markets which fulfill the requirements of all participants.

7.10 Fritz Henglein (Univ. of Copenhagen, DK & Deon Digital – Zürich, CH)

Fritz Henglein is Professor of Programming Languages and Systems at DIKU, the Department of Computer Science at the University of Copenhagen (UCPH) and Head of Research at Deon Digital AG, a Zürich/Copenhagen-based start-up developing secure and scalable digital contract technology for both decentralized (blockchain, distributed ledger) and centralized systems. His research interests and contributions are in semantic, logical and algorithmic aspects of programming languages, functional programming, domain-specific languages, digital contracts, reporting and analytics, smart contracts and distributed ledger technology, with applications in enterprise systems, business processes, high-performance computing, probabilistic programming and decentralized systems, including blockchain and distributed ledger systems.

Key desiderata of (next-generation) distributed ledger (DL) systems are tamper-proof, privacy-respecting recording of real-world and business events together with their evidence; secure distributed storage and transfer of assets such as money, assets, securities and (digital twins of) physical resources; automatically managed and enforced contracts that provide an effective and auditable basis for privacy preserving collaboration, analytics and planning; a balance of availability, consistency and network failure tolerance tuned to specific use case characteristics; and, most importantly, organizational decentralization to minimize the need for and competitive advantage of a dominant and controlling platform provider. Such a DL system facilitates tracking of digital resources, including valuable data, across organizations as well as transfer (and revocation) of control over them. Ongoing research, development and commercial application (at Deon Digital) of smart digital contract technology on existing commercial distributed ledger systems indicates that this may facilitate effective cross-organizational data exchange by facilitating unforgeable auditable proofs of authorized data use.

7.11 Matthias Jarke (RWTH Aachen University and Fraunhofer FIT, DE)

Data-driven machine learning methods are typically most successful when they can rely on very large and in some sense homogeneous training sets in areas where little prior scientific knowledge exists. Production engineering, management, and usage satisfy few of these criteria and therefore do not show many success stories, beyond narrowly defined specific issues in specific contexts. In contrast, the last years have seen impressive successes in model-driven materials and production engineering methods, these methods lack context and real-time adaptivity.

Our vision of an Internet of Production, pursued in an interdisciplinary DFG-funded Excellence Cluster at RWTH Aachen University, addresses these shortcomings: Through sophisticated heterogeneous data integration and controlled data sharing approaches, it broadens the experience base of cross-organizational product and process data. At the method level, it interleaves fast “reduced models” from different engineering fields, with enhanced explainable machine learning techniques and model-driven re-engineering during operations.

As a common conceptual modeling abstraction, we investigate Digital Shadows, a strongly empowered variant of the well-known view concept from data management. The idea of Digital Shadows dates back to Platon’s famour Cave Allegory but which he illustrated the limitations of all human knowledge – we can always only see partial perspectives on the world. Fifty years of data management research confirm that the growth of data has always outpaced our ability to deal with them, and we expect it to stay this way. Therefore, we see strong limitations for the currently fashionable Digital Twins when real-time and large scope are relevant, and propose to circumscribe them by well-structured collections of Digital Shadows. Besides enabling real-time monitoring and control at an abstract level, Digital Shadows are also well-suited for bridging interdisciplinary boundaries and communication problems between research and practice. Several initial experiments indicate the power of this approach but also highlight many further research challenges.

7.12 Jan Jürjens (Universität Koblenz-Landau, DE)

AI uses scientific methods, processes, algorithms, and systems to gain knowledge and insights into data that exists in a variety of formats. Characteristic of the area is on the one hand the high significance that it has. Based on the results obtained, numerous important decisions are made concerning the individual or society as a whole: diagnoses, therapies, credit decisions, spatial planning, etc. On the other hand, AI is characterized by the iterative and empirical-heuristic approach by which knowledge is extracted and decisions are derived. From the point of view of the provider of the data on which the AI-based analysis is performed, it is important that the Data Sovereignty is preserved in the context of the data analysis, which is closely related to the following aspects:

- Data security
- Data privacy
- Transparency and explainability
- Fairness / non-discrimination

Unfortunately, it is a significant challenge to be able to demonstrate whether or not these aspects of Data Sovereignty are satisfied in a given situation involving AI based analysis. In fact, it is already a challenge to just describe "correct behavior" of an AI based system, because its results are usually not predetermined and can only be obtained through the data analysis process, so it is in general not clear what "correctness" means in this context. The challenge is also that the behaviour of a self-learning algorithm depends on the training data to which it is continuously being subjected and therefore cannot be determined by only considering the algorithm itself. We thus need a verification approach which can be efficiently parameterized over the possible effects of the self-learning process without having to re-do the complete verification whenever the self-learning leads to a change in behaviour. The proposed talk discusses these challenges and presents an approach that supports the analysis of Data Sovereignty aspects in the context of AI-based systems based on the tool-based analysis of software design models in UML, which supports change-based verification and can thus deal with the different variants of algorithm behaviour arising from self-learning. The talk is based on work done as part of University of Koblenz' research priority programme "Engineering Trustworthy Data-intensive Systems" as well as within the context of the "International Data Spaces" initiative at Fraunhofer ISST.

7.13 Maurizio Lenzerini (Sapienza University of Rome, IT)

Data interoperability refers to the issue of accessing and processing data from multiple sources in order to create more holistic and contextual information for improving data analysis, for better decision-making, and for accountability purposes. In the era towards a data-driven society, the notion of data interoperability is of paramount importance. Looking at the reseach work in the last decades, several types of data interoperability scenarios emerged, including the following.

1. In Data Integration, we have a multitude of information sources, and we want to access them by means of a global schema, that somehow accomodates an integrated view of all data at the sources [35, 75].
2. In Data Exchange, we have a source databas, and a target database, and we want to move the data from the source to the target according to some specified criteria [5, 66].

3. In P2P Data Coordination, we have a network of information nodes (peers), and we want to let them communicate to each other in order to exchange data or queries [20, 76].
4. In Ontology-Based Data Management (OBDM), we have a collection of data sources and an ontol- ogy representing a semantic model of the domain of interest, and we want to govern (i.e., query, update, monitoring, etc.) the data at the sources through the ontology, rather than by interacting directly with the sources [29, 76].

A fundamental component of all the above data interoperability frameworks is the mapping. Indeed, put in an abstract way, all the above scenarios are characterized by an architecture constituted by various autonomous nodes (called databases, data sources, peers, etc.) which hold information, and which are linked to other nodes by means of mappings. A mapping is a statement specifying that some relationship exists between pieces of information held by one node and pieces of information held by another node. Specifically, in Data Integration the mappings relate the data sources to the global schema, in Data Exchange they relate the source database to the target database, in P2P Coordination they relate the various peers in the network, and in OBDM they relate the various data sources to the ontology. In the last years, many papers investigate the notion of mapping, from various points of view, and with different goals (see [67] and references therein). By looking at these papers, one could argue that one of the most important role of mapping is to allow reformulating queries expressed over a node into queries expressed over other mapped nodes. Such reformulation task is crucial, for example, for answering queries expressed over the global schema in a data integration system. Indeed, to compute the answer, the system has to figure out which queries to ask to the data sources (where the real data are located), and this is done by a step that we call direct rewriting: rewrite the query over the global schema in terms of a query over the data sources. A similar task has been studied in the other data interoperability scenarios. In OBDM, for instance, given a user queries expressed over the ontology, the aim is to find a direct rewriting of the query, i.e., a query over the source schema, that, once executed over the data, provides the user query answers that are logically implied by the ontology and the mapping. While the notion of direct rewriting has been the subject of many investigations in data interoperability in the last decades, in this paper we aim at discuss also a new notion of rewriting, that we call inverse rewriting. The importance of this new notion emerges when we consider the following task in the OBDM scenario: Given a query q over the sources, find the query over the ontology that characterizes q at best (independently from the current source database). Note that the problem is reversed with respect to the one where the traditional (direct) rewriting is used: here, we start with a source query, and we aim at deriving a corresponding query over the ontology. Thus, we are dealing with a sort of reverse engineering problem, which is novel in the investigation of data interoperabilty. We argue that this problem is relevant in a plethora of application scenarios. For the sake of brevity, we mention only three of them. (1) Following the ideas in [25], the notion of reverse rewriting can be used to provide the semantics of open data and open APIs published by organizations, which is a crucial aspect for unchaining all the potentials of open data. (2) Although the architecture of many modern Information Systems is based on data services, that are abstractions of computation done on data sources, it is often the case that the semantics of such computations is not well specified or documented. Can we automatically produce a semantic characterization of a data service, having an OBDA specification available? The idea is to exploit a new reasoning task over the OBDA specification, that works as follows: we express the data service in terms of a query over the sources, and we use the notion of reverse rewriting for deriving the query over the ontology that best describes the data service, given the ontology and the mapping.

(3) It can be shown that the concept of reverse rewriting is also useful for a semantic-based approach to source profiling [2], in particular for describing the structure and the content of a data source in terms of the business vocabulary.

7.14 Wolfgang Maaß (Universität des Saarlandes – Saarbrücken, DE)

Intelligent services using Artificial Intelligence methods are tools for creating, modifying and merging data products from different industries, in particular industrial production systems (Industrie 4.0). We take an extended view on data products, which includes (a) data as statements about a domain, (b) software, and (c) models including conceptual models, machine learning models, and semantic models. In our research we develop AI algorithms and architectures for distributed (IOT) environments. These technologies are used to build domain-specific smart service systems. For example, Smart Dialog Services are used to make data products and derive decision recommendations based on machine learning mechanisms accessible to human experts (explainable AI and responsible AI). A further focus is the investigation of means for the automatic evaluation of data products as input for the financial reporting of digital assets. Our research has been conducted in various fields, such as industrial manufacturing, health care/medicine and media. While the media industry and medicine have been struggling with the challenges of digital transformation for about 30 years, industrial production is at the beginning of understanding digital products as an independent economic asset class.

7.15 Paolo Missier (Newcastle University, GB)

Data marketplaces are becoming ubiquitous, but for the most part, they assume (i) static data, and (ii) a trusted environment where data trading takes place. Both these assumptions introduce limitations in the potential of individuals to exchange their own personal data with other parties. Data streams that originate from Internet of Things (IoT) devices, often placed in people's premises (house, vehicle, or own body using wearables), are increasingly viewed as tradeable assets with value not only to the device owners, but also with resell value, i.e., to third party buyers. We envision a marketplace for IoT data streams that can unlock such potential value in a scalable way, by enabling any pairs of data providers and consumers to engage in data exchange transactions without any prior assumption of mutual trust. We have recently proposed one such next-generation marketplace model, where the requirement for trust in data exchange is fulfilled using blockchain technology and specifically Smart Contracts. I am interested in discussing such marketplace model within the context of this Dagstuhl Seminar, as I believe it has the potential to become a cornerstone of data ecosystem where organisations exchange their data reliably and in real-time, subject to legally binding trading agreements, and while providing incentives for fair trading.

7.16 Boris Otto (Fraunhofer ISST – Dortmund, DE & TU Dortmund, DE)

The proliferation of digital technologies such as cloud platforms and cyber-physical systems in the manufacturing industry enables process and service innovation in production and supply networks. Sharing data among network partners in an important prerequisite for leveraing this innovation potential. In this context, the notion of "digital twins" has received significant attention, both in the scientific and the practioners' community. In general, a digital twin represents real world objects (such as production equipment, components, material etc.) along their lifecycle. It consists of descriptive data (e.g. the eCl@ss number of a tool machine component) as well as event data (such as temperature, vibration data etc. captured in the production process). Furthermore, the digital twin comprises both metadata and data items. A prominent digital twin information model, namely the so-called Asset Administraive Shell proposed by the German "Plattform Industrie 4.0" distinguihes between "types" and "instances". Recent digital twin scenarios are based sharing digital twin data within a production and supply network or even within an industrial ecosystem. Hence, the concept of "shared digital twins" of components, machines etc. emerges. While the fundamental concepts of digital twins in general (such as metadata, data models, distributed data storage are mature research topics, sharing digital twin data among different partners has not sufficentily be conceptualized. Moreover, taking an Information System Research perspective, a set of interesting research questions can be identified. Examples are:

- What are appropriate query approaches to retrieve shared digital twin data?
- How to distribute storing and processing of shared digital twin data between edge, fog, and cloud level? What are appropriate synchronization approaches?
- How do ensure data sovereignty of different contributers to the shared digital twin in a complex production and supply network?
- How to design functional data governance models for shared digital twin?
- How to enable access and usage control for shared digital twin data?
- How to design data provenance architectures for shared digital twins.

Based on a solid conceptulization of shared digital twins, a research agenda is required to identify research demand and outline promising research trajectories.

7.17 Elda Paja (IT University of Copenhagen, DK)

My research focuses on techniques to support the elicitation or design of socio-technical requirements (looking not only at software, but also humans and organizations, their interactions and message exchanges), requirements that come from different stakeholders, and might be conflicting with one another or impact business policies or system functionality. In my work, I have been exploring the design of socio-technical systems from different angles: security, privacy, risk, and decision-making support. The common ground of my approach is the use of conceptual modelling techniques, and in particular the use of modelling primitives to describe the rationale behind the behaviour of the various participants of a socio-technical system, and the definition and exploitation of automated reasoning techniques to support the work of requirements engineers. Contributions to the seminar

- Discussing on analogies and potential of adapting existing methods for requirements engineering for data ecosystems, treating data as first class citizens.
- Presenting recent work on consent privacy requirements for sociotechnical systems, following a multi-level approach, starting from social and organizational requirements, to business process level verification of deviations or breaches.
- Presenting recent work on consent verification monitoring, namely a formal framework to support companies and users in their understanding of policies evolution under consent regime that supports both retroactive and non-retroactive consent and consent revocation, all in a context where personal data provides important business value, e.g. in the personalization of services.

7.18 Barbara Pernici (Polytechnic University of Milan, IT)

Sharing of scientific and scholarly data is enabled by open or shared repositories in many different scientific domains. Data sharing and open data are not final goals in themselves, however, and the real benefit is in data reuse by different actors.

Focusing on reuse, the design of integrated frameworks that make it simple and effective both the upload and the retrieval of large amounts of scientific data extracted from the literature or produced by research labs, to support research, and in particular scientific model development, is a challenging issue. In particular the development of scientific models to reproduce and predict complex phenomena is a challenging task, which requires a rich set of data both for model development and validation. This task is becoming particularly critical as more and more approaches to automate model development using machine learning techniques emerge in different fields.

Starting from ongoing researches in the domain of chemistry engineering and emergency information systems, several research questions emerge and will be discussed:

- How to represent metadata in a flexible way, in particular when there is no agreement within the ecosystem of contributors on a common underlying model and new features emerge during the analysis.
- How to allow several organizations simultaneously analyze the data from different perspectives and integrate their results.
- How to assess the quality of the data and of the analysis/classification models, possibly automatically derived through deep learning mechanisms, distinguishing between assessment of quality of data (both training and input data) and other quality issues introduced by the analysis models.
- Another open issue concerns the intellectual property and access to results both of the original data and of data obtained through modeling/ simulation processes.
- Finally, the interaction of information systems experts, computer scientists and experts on the specific domain of interest requires new flexible approaches and design patterns in the development and of such systems.

7.19 Frank Piller (RWTH Aachen, DE)

One of our core research streams at the Institute for Technology and Innovation Management at RWTH Aachen University investigates the need of established corporations to deal with disruptive business model innovation and supporting organizational structures and cultures. The rise of platform-based business models (or business models for industrial data ecosystems) is one of the main drivers of change in this field. Hence, we currently have a number of research projects where we study the systematic development of these platform-based business ecosystems and the strategic positioning of an industrial company in these ecosystems. The largest among these projects is the Cluster of Excellence "Internet of Production (IoP)", funded by the German Research Council (DFG), Project ID 390 621 612). The IoP resembles the vision of an open network of sensors, assets, products, and actors that continuously generate data. A core element hence is the (re-)use of data, digital shadows, and applications by other parties to facilitate faster and more efficient learning and analytics. The rise of platforms (business ecosystems) where these data are being exchanged and enhanced by dedicated "apps" is a central element of the IoP vision. To create value, ecosystems build on complementary inputs made by loosely interconnected, yet independent stakeholders. Among these participants, dedicated mechanisms governing data access and privacy are required. Our research here takes an inter-firm (external) perspective: setting the right incentives for sharing deep production know-how and data while balancing value creation with value capture (sharing the rents) for all participants. Among others, we are interested in the following research questions:

- Modeling the tension between openness in value creation and control of value capture,
- Managing property rights (access, transfer, enforcement) at data, applications, and connected assets as a result of varying degrees of platform openness, and
- Definition of governance modes and design factors to generate adequate business models for the IoP that allow to maximize value appropriation for all involved actors.

7.20 Andreas Rausch (TU Clausthal, DE)

Data today plays a much bigger role than it used and now data is the new oil. The reason why data is generated and collected in such a massive amount is pretty simple. Data is the new oil and has become a fuel for new business models. However, most of the available data is related to very limited number of companies and organizations that form almost an oligopoly – the so-called GAFA companies [18] [19]: Google, Facebook, Apple and Amazon. For Small and Medium Enterprises (SMEs) it is much harder to extract the same value from their data compared to these large enterprises. On one hand it is very difficult for enterprises (Data consumers) to obtain proper data and on other hand for those who collect data (Data providers), the problem is- How to draw to additional profit from the data beyond its obvious purpose. Thus, a common data sharing platform is required where the data producers can obtain profit from their data and the data consumers can easily find data. To tackle this a new data marketplace ecosystem is required based on just technical aspects but also the social and economic aspects. Rather than conventional centralized solutions, our research focuses on solutions for enabling the data sharing without the stakeholders having to have full trust in the marketplace owner or provide.

The three main building blocks on this new proposed “Data marketplace ecosystem” are a community system, open business architecture platform and the relationship between community system and open business architecture platform. This data marketplace ecosystem is a decentralized, open and large software system, which is owned, controlled and used by a community system.

- A community system: A community system is a group of people who share a common interest but still form a heterogeneous system. The community system can be subdivided into different homogeneous subgroups.
  - Provide community- The provider community defines some rules and standards for the ecosystem to function safely. It also helps the ecosystem evolve based on the requirements of the user community.

[18] Z. Abrahamson, “Essential Data,” 2014.

[19] Lucy Handley, “Amazon beats Apple and Google to become the world’s most valuable brand,” 2019. [Online]. Available: https://www.cnbc.com/2019/06/11/amazon-beats-apple-and-google-to-become-the-worlds-most-valuable-brand.html [Accessed: 18-Jun-2019]

  - User community: The users of the data marketplace are the data providers and the consumers.
  - Operator community: The operator community operates the data marketplace ecosystem. The operating community provides the computation and technical infrastructure for the data marketplace to function. The goal to have such a community avoid single ownerships.
- Open business architecture platform: The open business architecture platform is the platform i.e. the data marketplace itself which the community systems develop, operate and use. It describes the technical realization of a whole system for a data marketplace ecosystem. The main goal here is to provide a completely decentralized data marketplace ecosystem, which is open and flexible as far as possible but still provides nonfunctional requirements like safety, security, privacy and dependability.
- Relationship between the community system and open business architecture platform: The community system has various responsibilities which helps defining, development and the evolvement of the open business architecture platform. These responsibilities are the relationship between the community system and open business architecture platform. As the community uses and evolves the ecosystem understanding this relation is very important.

The proposed concept is community driven and proposes on the one side an initial concept for the community structure and on the other side an architecture which is open, flexible and secure and is aligned to this community. Nevertheless, this project is still work in progress and will be continuously expanded in the near future.

Although such a data ecosystem provides new opportunities for data consumers, providers and many other stakeholders, there are many challenges to be tackled. We identified various challenges in the data marketplace ecosystems. Although while writing this, our research is still work in progress we sketch some solutions for tackling few challenges.

- Open and secure infrastructure: In order to provide a fairness and transparency, new methods for secure but an open infrastructure is required.
- Metadata: In order to sell the data on the marketplace the seller needs to provide some description about the data. E.g. What is the data about, the size of the data etc. This information about the data is known as meta data. On one hand meta data gives the buyer the data information and is also used for search which helps the buyers to find relevant data. But providing such detailed relevant information about the data can be too much work for the sellers. One possibility for this challenge is Automatic generation of metadata. But the question still is how to generate the meta data automatically and protect it from unauthorized access.
- Data quality: Data is very different the any other commodity sold on electronic marketplaces. When a user buys a physical product online, this product can be returned or exchanged if he/she is unsatisfied with it. But the same does not applies for datasets. Once the buyer sees the datasets i.e. buys them, it cannot be returned. Thus, the biggest challenge we identify for trading data as a commodity is ensuring the data quality.

7.21 Jakob Rehof (TU Dortmund, DE)

When data ecosystems are understood to involve distributed networks of data providers and data consumers the question arises how to organize data logistics in the sense of a data supply network. Such issues may reasonably be attributed to the sub-field of data engineering.

A central data engineering challenge in this context is the automation of data supply networks, that is, essentially, the task of getting data from A to B in a form understood and desired by A and B. If two or more parties wanting to share data need to engage in arbitrarily complicated software engineering projects prior to making data flow among themselves, the prospect of realizing data ecosystems at large is fundamentally inhibited.

An interesting line of research motivated by these observations concerns the employment of component-oriented synthesis (much in the sense discussed at the previous Dagstuhl Seminar 14232 on Design and Synthesis from Components) as a means towards automating data supply chains. In particular, the automatic generation of complex data transformation functions and rendering functions appears to be a useful goal which should not be too far out of reach for practical exploitation. Based on some years of experience with research in type-based component-oriented synthesis in Dortmund, the idea of developing specific frameworks for use in data supply networks appears natural. The capacity of such frameworks to operate relative to given (but possibly dynamically changing) repositories of components seems useful here. For example, data provider A could inject data transformation functions based on domain specific knowledge of data semantics together with accompanying metadata. Data consumer B could inject domain specific transformers and rendering functions pertaining to the domain of use. An interesting fact about such a scenario is that the repository could evolve in a distributed fashion without any need for a central control other than the logical control implicit in overarching standards such as may be embodied in metadata formats. Thus a vision may arise which foresees distributed and (to a large extent) self-organizing networks of functionalities (the repository) accessible to automatic methods of code generation (synthesis) composing functions on demand to achieve a stated goal (for example, the transformation of data from A to B in a desired form). One can go further and imagine such a framework being combined with query languages and query mechanisms tailored for the network (such that, e.g., a data transformation function is automatically synthesized as a side-effect of executing a query against the network).

7.22 Simon Scerri (Fraunhofer IAIS – Sankt Augustin, DE)

Due to the well-understood AI opportunities presenting themselves in the last few years, there has been a steadily increasing and consistent interest in data sharing methods, solutions and practices. This has been observed internationally both at an industrial level, as well as at a political level. Seeing this as an opportunity to boost the data economy in Europe, the European Commission has reacted strongly by organising many events at attempting to invest in the convergence of technology and infrastructures that can enable the realisation and adoption of a pan-European data sharing space that can incorporate existing vertical, cross-sectoral, personal and industrial data spaces and enable broader participation. The realisation of an ’open’ data sharing ecosystem that can serve the needs of all kinds of stakeholders also introduces exciting opportunities for scenarios that are not only restricted to B2B, but also enable data exchange possibilities for science, government and private citizens. For this vision to be achieved, the convergence of efforts at both industrial and socio-political level, through the adoption of standards that respect the existing legal and regulatory frameworks (e.g. GDPR for personal data), is being encouraged.

In view of this European vision for a data sharing space, the Big Data Value Association, an industry-driven international not–for-profit organisation with 200 European members (composed of large, small, and medium-sized industries as well as research and user organizations) has an activity group following and promoting advances in Data Sharing ecosystems.

Simon Scerri is one of the lead editors for the position paper published on the topic [20]. In its ambition to support the convergence of existing efforts, standards and technology in this sphere, the group is continuously seeking to collaborate with external experts and data practitioners. The position paper, for which a second version is planned for early 2020, includes a survey of the broad (international) technical landscape, and enlists the known opportunities (for business, science, government and public bodies and citizens), known challenges (legal compliance, technical, business and organisational, national and regional) and a list of key recommendations that can pave the way for the targeted convergence to materialise. The recommendations are addressed to both European policy makers and industry alike, as the two primary entities identified as having the highest likelihood of accelerating advances in the area.

The technical concerns and requirements discussed in the Dagstuhl seminar have helped to both confirm the challenges identified, as well as to bring to the focus additional concerns and relevant initiatives. In particular, the 'blueprints' for a high-level generic data sharing architecture that is open to all will be considered for a similar illustration in the second version of the BDVA position paper. For further information or updates, please contact Simon or refer to the BDVA Website Downloads section in the near future. In addition, an appeal is made for entities with an interest in data sharing practices to refer to a survey which is due to be published and promoted by the BDVA in January.

7.23 Julian Schütte (Fraunhofer AISEC – München, DE)

Future data ecosystems go beyond centralized cloud services and will require new technical paradigms and infrastructures to establish trust among participants. Central cloud providers currently serve not only as Identity Providers (IdP), but also as implicitly trusted data brokers which are expected to treat sensitive business data confidential, reliably enforce access constraints, and provide accounting and billing services. However, many upcoming business cases cannot be served by a single trusted cloud service – either due to legal reasons or due to technical necessities that require an orchestration of several services and stakeholders.

To overcome the limitations imposed by central cloud services, technical infrastructures that allow direct data exchange between participants without relying on central cloud services are needed. This raises various research questions, such as the establishment of trust between peers, the control of data flowing between services across enterprise boundaries, and ways to protect data while still being able to process it.

While some of these individual problems have been addressed by the research community in the past, synergies from recent developments in distributed architectures, in advanced cryptography, and in upcoming commodity hardware will allow to create a technical foundation for data-driven business cases that do not depend on centralized cloud providers. I would like to make the following contributions:

- Discussion of the main security building blocks needed to create trusted decentralized data ecosystems
- Insights into data usage control systems and their technical manifestation
- Presentation and discussion of mechanisms and upcoming standards for trust establishment, i.e. the foundations of automatically establishing trust in a remote party at a technical level

[20] http://www.bdva.eu/sites/default/files/BDVA%20DataSharingSpace%20PositionPaper_April2019_V1.pdf

- Examples on how novel security mechanisms such as cryptographic access control in decentralized ledgers will foster new business cases in the logistics domain

7.24 Egbert Jan Sol (TNO – Eindhoven, NL)

The evolution of our (mechanical/electronics) industry towards a data driven industry (data eco-systems) involves many technologies. IoT data collection is one of the many aspects. It is the basis for coupling data to Digital Twin, AI-algorithm, but also data-driven businesses as servitization where measured usage of products leads to e.g. billing. IoT data collection faces several challenges to overcome. There are far too many different (fieldbus type) IoT data communication standards, limiting data use to local usage. But standardization as OPC-UA and usage of 5G will lead to an explosion of (large/big) data sets within and across businesses. Collected IoT data is not copy-right protected as it doesn't involved creative labor and needs extra legal and cyber security measurement. For AI usage data must be cleaned and for sending bills certain IoT data must be treated with DLT (distributed ledger tech/blockchain tech). And, maybe the biggest problem for business, is a huge lack of skilled, trained people with the proper digital skills as practically every non-university trained person above 35 years today didn't get any training in digital technologies 20 years ago at school when they were 15 years. With the rapid deployment of Industrie 4.0 mankind faces for the first time in history the need to life-long retrain every one in digital skills. For me the main question is what are the new big data/AI etc data technologies that are needed, that will be developed and how to make them usable, cq how can we educate and train people (and politicians) at academic, higher level and vocational level to understand them, to create the proper data-ecosystems and to deploy them. This challenge is not limited to my background in manufacturing industries, but is similar for food, medical and many other domains.

7.25 Maria-Esther Vidal (TIB – Hannover, DE)

During my academic and profesional work, I have been involved in diverse projects where the resolution of interoperability and quality issues across large volumes of heterogeneous data sources is a pre-condition for effectively devising data integration. Albeit being projects from different areas, e.g., industry, biomedical, or scholarly communication, data variety and veracity seemed to be domain-agnostic even thought their resolution required domain specific knowledge. Motivated by this fact, the definion of generic frameworks able to identify interoperability issues while allowing for an effective data integration process has been one of my research topics and one of my interests in attending the seminar on Data Ecosystems. Currently, I am leading the tasks of integrating clinical data from electronic health records, gene sequences, and medical images, with open pharmacogenomics data and scientific publications. Natural language processing and semantics annotations from controlled vocabularies provide the basis for the fusion of these variety of data sources. During the seminar, I can contribute describing both challenges and solutions that we have tackled in the context of projects like iASiS [21] and BigMedilytics [22]. I can also discuss, the approaches that we have considered to address these issues during query processing.

[21] http://project-iasis.eu/

[22] https://www.bigmedilytics.eu/

References

1 Behzad Abdolmaleki, Karim Baghery, Helger Lipmaa, and Michał Zając. A subversion-resistant SNARK. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

2 Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Data profiling: A tutorial. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 1747–1751, 2017.

3 Ron Adner. Match your innovation strategy to your innovation ecosystem. *Harvard business review*, 84(4):98, 2006.

4 Dimitris Apostolou, Gregoris Mentzas, Bertin Klein, Andreas Abecker, and Wolfgang Maass. Interorganizational knowledge exchanges. *IEEE Intelligent Systems*, 23(4):65–74, 2008.

5 Marcelo Arenas, Pablo Barceló, Leonid Libkin, and Filip Murlak. *Foundations of Data Exchange*. Cambridge University Press, 2014.

6 Carlo Batini and Monica Scannapieco. *Data and Information Quality - Dimensions, Principles and Techniques*. Data-Centric Systems and Applications. Springer, 2016.

7 Eli Ben-Sasson, Iddo Bentov, Yinon Horesh, and Michael Riabzev. Scalable, transparent, and post-quantum secure computational integrity. *Eprint.Iacr.Org*, 2018.

8 Eli Ben-Sasson, Alessandro Chiesa, Eran Tromer, and Madars Virza. Succinct Non-Interactive Zero Knowledge for a von Neumann Architecture. *USENIX Security Symposium*, 2014.

9 Alexander Benlian, Daniel Hilkert, and Thomas Hess. How open is this platform? the meaning and measurement of platform openness from the complementers' perspective. *Journal of Information Technology*, 30(3):209–228, 2015.

10 Jan Bessai, Tzu-Chun Chen, Andrej Dudenhefner, Boris Düdder, Ugo de'Liguoro, and Jakob Rehof. Mixin composition synthesis based on intersection types. *Logical Methods in Computer Science*, 14(1), 2018.

11 Jan Bessai, Andrej Dudenhefner, Tzu Chun Chen, Ugo DE'LIGUORO, Jakob Rehof, et al. Mixin composition synthesis based on intersection types. In *TLCA 2015*, volume 38, pages 76–91. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015.

12 Jan Bessai, Andrej Dudenhefner, Boris Düdder, Moritz Martens, and Jakob Rehof. Combinatory logic synthesizer. In *International Symposium On Leveraging Applications of Formal Methods, Verification and Validation*, pages 26–40. Springer, 2014.

13 Nitesh Bharosa, Marijn Janssen, Bram Klievink, and Yao-hua Tan. Developing multi-sided platforms for public-private information sharing. In Sehl Mellouli, Luis F. Luna-Reyes, and Jing Zhang, editors, *Proceedings of the 14th Annual International Conference on Digital Government Research - dg.o '13*, page 146, New York, New York, USA, 2013. ACM Press.

14 Vittorio Dal Bianco, Varvana Myllarniemi, Marko Komssi, and Mikko Raatikainen. The role of platform boundary resources in software ecosystems: A case study. In *2014 IEEE/IFIP Conference on Software Architecture*, pages 11–20. IEEE, 2014.

15 Kevin Boudreau. Open platform strategies and innovation: Granting access vs. devolving control. *Management science*, 56(10):1849–1872, 2010.

16 Sean Bowe, Alessandro Chiesa, Matthew Green, Ian Miers, Pratyush Mishra, Howard Wu, Cornell Tech, and Howard Wu. Zexe: Enabling Decentralized Private Computation. *Cryptology ePrint Archive*, 2018.

17 Georg Bramm, Mark Gall, and Julian Schütte. BDABE - blockchain-based distributed attribute based encryption. In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, ICETE 2018 - Volume 2: SECRYPT, Porto, Portugal, July 26-28, 2018*, pages 265–276, 2018.

18 Benedikt Bünz, Jonathan Bootle, Dan Boneh, Andrew Poelstra, Pieter Wuille, and Greg Maxwell. Bulletproofs: Short Proofs for Confidential Transactions and More. In *Proceedings - IEEE Symposium on Security and Privacy*, 2018.

19 Christian Burmeister, Dirk Lüttgens, and Frank T Piller. Business model innovation for industrie 4.0: why the "industrial internet" mandates a new perspective on innovation. *Die Unternehmung*, 70(2):124–152, 2016.

20 Diego Calvanese, Elio Damaggio, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. Semantic data integration in P2P systems. In *Databases, Information Systems, and Peer-to-Peer Computing, First International Workshop, DBISP2P, Berlin Germany, September 7-8, 2003, Revised Papers*, pages 77–90, 2003.

21 Ran Canetti, Yuval Ishai, Ravi Kumar, Michael K Reiter, Ronitt Rubinfeld, and Rebecca N Wright. Selective private function evaluation with applications to private statistics. In *Proceedings of the twentieth annual ACM symposium on Principles of distributed computing*, pages 293–304. ACM, 2001.

22 Michelle Cheatham, Isabel F. Cruz, Jérôme Euzenat, and Catia Pesquita. Special issue on ontology and linked data matching. *Semantic Web*, 8(2):183–184, 2017.

23 Hsinchun Chen, Roger HL Chiang, and Veda C Storey. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4), 2012.

24 Soon-Yong Choi, Dale O Stahl, and Andrew B Whinston. *The economics of electronic commerce.* Macmillan Technical Publ. Indianapolis, 1997.

25 Gianluca Cima. Preliminary results on ontology-based open data publishing. In *Proceedings of the 30th International Workshop on Description Logics, Montpellier, France, July 18-21, 2017*, 2017.

26 George Coker, Joshua Guttman, Peter Loscocco, Amy Herzog, Jonathan Millen, Brian O'Hanlon, John Ramsdell, Ariel Segall, Justin Sheehy, and Brian Sniffen. Principles of remote attestation. *International Journal of Information Security*, 10(2):63–81, 2011.

27 Diego Collarana, Mikhail Galkin, Ignacio Traverso-Ribón, Maria-Esther Vidal, Christoph Lange, and Sören Auer. MINTE: semantically integrating RDF graphs. In *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, WIMS*, 2017.

28 Brice Dattée, Oliver Alexy, and Erkko Autio. Maneuvering in poor visibility: How firms play the ecosystem game when uncertainty is high. *Academy of Management Journal*, 61(2):466–498, 2018.

29 Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, and Riccardo Rosati. Using ontologies for semantic data integration. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, pages 187–202. Springer, 2018.

30 Mark de Reuver, Bouke Nederstigt, and Marijn Janssen. Launch strategies for multi-sided data analytics platforms. In *ECIS 2018*, 2018.

31 Yuri Demchenko, Cees de Laat, and Peter Membrey. Defining architecture components of the big data ecosystem. In *2014 International Conference on Collaboration Technologies and Systems (CTS)*, pages 104–112. IEEE, 2014.

32 Yuri Demchenko, Wouter Los, and Cees Laat. Data as economic goods: Definitions, properties, challenges, enabling technologies for future data markets. *ITUJournal - ICT Discoveries*, 1(2), 2018.

33 Charles Dhanaraj and Arvind Parkhe. Orchestrating innovation networks. *Academy of management review*, 31(3):659–669, 2006.

34 AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. *Principles of Data Integration.* Morgan Kaufmann, 2012.

35 AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. *Principles of Data Integration.* Morgan Kaufmann, 2012.

36 Boris Düdder and Omri Ross. Timber tracking: Reducing complexity of due diligence by using blockchain technology. *Available at SSRN 3015219*, 2017.

37 Jennie Duggan, Aaron J. Elmore, Michael Stonebraker, Magda Balazinska, Bill Howe, Jeremy Kepner, Sam Madden, David Maier, Tim Mattson, and Stan Zdonik. The BigDAWG Polystore System. *SIGMOD Rec.*, 44(2):11–16, August 2015.

38 Ben Eaton, Silvia Elaluf-Calderwood, Carsten Sørensen, and Youngjin Yoo. Distributed tuning of boundary resources: the case of apple's ios service system. *MIS Quarterly*, 39(1):217–243, 2015.

39 Benjamin Egelund-Müller, Martin Elsman, Fritz Henglein, and Omri Ross. Automated execution of financial contracts on blockchains. *Business & Information Systems Engineering*, 59(6):457–467, 2017.

40 Kemele M. Endris, Philipp D. Rohde, Maria-Esther Vidal, and Sören Auer. Ontario: Federated query processing against a semantic data lake. In *Database and Expert Systems Applications - 30th International Conference, DEXA 2019, Linz, Austria, August 26-29, 2019, Proceedings, Part I*, pages 379–395, 2019.

41 Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching, Second Edition*. Springer, 2013.

42 Ronald Fagin, Laura M. Haas, Mauricio A. Hernández, Renée J. Miller, Lucian Popa, and Yannis Velegrakis. Clio: Schema mapping creation and data exchange. In *Conceptual Modeling: Foundations and Applications - Essays in Honor of John Mylopoulos*, pages 198–236, 2009.

43 Michael Franklin, Alon Halevy, and David Maier. From databases to dataspaces. *ACM SIGMOD Record*, 34(4):27–33, 2005.

44 Avigdor Gal. Uncertain schema matching. In *Encyclopedia of Big Data Technologies*. Springer, 2019.

45 Avigdor Gal, Haggai Roitman, and Roee Shraga. Heterogeneous data integration by learning to rerank schema matches. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*, pages 959–964, 2018.

46 Günter Gans, Matthias Jarke, Stefanie Kethers, and Gerhard Lakemeyer. Continuous requirements management for organisation networks: a (dis)trust-based approach. *Requir. Eng.*, 8(1):4–22, 2003.

47 Günter Gans, Matthias Jarke, Gerhard Lakemeyer, and Dominik Schmitz. Deliberation in a metadata-based modeling and simulation environment for inter-organizational networks. *Information Systems*, 30(7):587–607, 2005.

48 Annabelle Gawer. Platform dynamics and strategies: from products to services. In Annabelle Gawer, editor, *Platforms, Markets and Innovation*, pages 45–77. Edward Elgar Publishing, 2009.

49 Annabelle Gawer. Bridging differing perspectives on technological platforms: Toward an integrative framework. *Research policy*, 43(7):1239–1249, 2014.

50 Craig Gentry. *A fully homomorphic encryption scheme*. PhD thesis, Stanford University Stanford, 2009.

51 Cheng Hian Goh, Stéphane Bressan, Stuart Madnick, and Michael Siegel. Context interchange: New features and formalisms for the intelligent integration of information. *ACM Trans. Inf. Syst.*, 17(3):270–293, 1999.

52 Behzad Golshan, Alon Y. Halevy, George A. Mihaila, and Wang-Chiew Tan. Data Integration: After the Teenage Years. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017*, pages 101–106, 2017.

53 Simon Gottschalk, Nicolas Tempelmeier, Günter Kniesel, Vasileios Iosifidis, Besnik Fetahu, and Elena Demidova. Simple-ml: Towards a framework for semantic data analytics workflows. In *Semantic Systems. The Power of AI and Knowledge Graphs - 15th International*

Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9-12, 2019, Proceedings, pages 359–366, 2019.

54 Rihan Hai, Sandra Geisler, and Christoph Quix. Constance: An intelligent data lake system. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 2097–2100, 2016.

55 Alon Y. Halevy, Anand Rajaraman, and Joann J. Ordille. Data Integration: The Teenage Years. In *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006*, pages 9–16, 2006.

56 George Heineman, Armend Hoxha, Boris Düdder, and Jakob Rehof. Towards migrating object-oriented frameworks to enable synthesis of product line members. In *Proceedings of the 19th International Conference on Software Product Line*, pages 56–60. ACM, 2015.

57 O. Henfridsson and B. Bygstad. The generative mechanisms of digital infrastructure evolution. *MIS Quarterly: Management Information Systems*, 37(3):907–931, 2013.

58 Fritz Henglein and Jakob Rehof. Modal intersection types, two-level languages, and staged synthesis. In *Semantics, Logics, and Calculi - Essays Dedicated to Hanne Riis Nielson and Flemming Nielson on the Occasion of Their 60th Birthdays*, volume 9560 of *Lecture Notes in Computer Science*, pages 289–312, 2016.

59 Anne Immonen, Marko Palviainen, and Eila Ovaska. Requirements of an open data based business ecosystem. *IEEE Access*, 2:88–103, 2014.

60 Matthias Jarke. Data spaces: combining goal-driven and data-driven approaches in community decision and negotiation support. In *International Conference on Group Decision and Negotiation*, pages 3–14. Springer, 2017.

61 Matthias Jarke, Manfred Jeusfeld, and Christoph Quix. Data-centric intelligent information integration–from concepts to automation. *Journal of Intelligent Information Systems*, 43(3):437–462, 2014.

62 Thorhildur Jetzek, Michel Avital, and Niels Bjørn-Andersen. Generating sustainable value from open data in a sharing society. In *International Working Conference on Transfer and Diffusion of IT*, pages 62–82. Springer, 2014.

63 Manfred A. Jeusfeld, Matthias Jarke, and John Mylopouos. *Metamodeling for Method Engineering*. MIT Press, 2010.

64 Yasar Khan, Antoine Zimmermann, Alokkumar Jha, Vijay Gadepally, Mathieu D'Aquin, and Ratnesh Sahay. One size does not fit all: Querying web polystores. *IEEE Access*, 7:9598–9617, 01 2019.

65 Craig A. Knoblock and Pedro A. Szekely. Exploiting Semantics for Big Data Integration. *AI Magazine*, 36(1):25–38, 2015.

66 Phokion G. Kolaitis. Schema mappings, data exchange, and metadata management. In *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 13-15, 2005, Baltimore, Maryland, USA*, pages 61–75, 2005.

67 Phokion G. Kolaitis. Reflections on schema mappings, data exchange, and metadata management. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Houston, TX, USA, June 10-15, 2018*, pages 107–109, 2018.

68 Constantine E. Kontokosta. Energy disclosure, market behavior, and the building data ecosystem. *Annals of the New York Academy of Sciences*, 1295:34–43, 2013.

69 Sebastian Kortmann, Carsten Gelhard, Carsten Zimmermann, and Frank T Piller. Linking strategic flexibility and operational efficiency: The mediating role of ambidextrous operational capabilities. *Journal of Operations Management*, 32(7-8):475–490, 2014.

70 Sebastian Kortmann and Frank Piller. Open business models and closed-loop value chains: Redefining the firm-consumer relationship. *California Management Review*, 58(3):88–108, 2016.

71 Anastasia Krithara, Fotis Aisopos, Vassiliki Rentoumi, Anastasios Nentidis, Konstantinos Bougiatiotis, Maria-Esther Vidal, Ernestina Menasalvas, Alejandro Rodríguez González, Eleftherios Samaras, Peter Garrard, Maria Torrente, Mariano Provencio Pulla, Nikos Dimakopoulos, Rui Mauricio, Jordi Rambla De Argila, Gian Gaetano Tartaglia, and George Paliouras. iasis: Towards heterogeneous big data analysis for personalized medicine. In *32nd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2019, Cordoba, Spain, June 5-7, 2019*, pages 106–111, 2019.

72 Alice LaPlante and Ben Sharma. *Architecting data lakes: data management architectures for advanced business use cases.* O'Reilly Media, Sebastopol, 2016.

73 Kristin Lauter, Michael Naehrig, and Vinod Vaikuntanathan. Can homomorphic encryption be practical? In *Proceedings of the ACM Conference on Computer and Communications Security*, 2011.

74 Maurizio Lenzerini. Data Integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, Madison, Wisconsin, USA*, pages 233–246, 2002.

75 Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, Madison, Wisconsin, USA*, pages 233–246, 2002.

76 Maurizio Lenzerini. Managing data through the lens of an ontology. *AI Magazine*, 39(2):65–74, 2018.

77 Xavier Leroy. A formally verified compiler back-end. *Journal of Automated Reasoning*, 43(4):363, 2009.

78 Marten Lohstroh and Edward A Lee. An interface theory for the internet of things. In *SEFM 2015 Collocated Workshops*, pages 20–34. Springer, 2015.

79 Wolfgang Maass. *Elektronische Wissensmärkte: Handel von Information und Wissen über digitale Netze.* Springer-Verlag, 2009.

80 Wolfgang Maass, Wernher Behrendt, and Aldo Gangemi. Trading digital information goods based on semantic technologies. *Journal of Theoretical and Applied Electronic Commerce Research*, 2(3):18–35, 2007.

81 Majid Mohammadi, Amir Ahooye Atashin, Wout Hofman, and Yao-Hua Tan. Comparison of Ontology Alignment Systems Across Single Matching Task Via the McNemar's Test. *TKDD*, 12(4):51:1–51:18, 2018.

82 Corrado Moiso and Roberto Minerva. Towards a user-centric personal data ecosystem: The role of the bank of individuals' data. In Stuart Sharrock, editor, *2012 16th International Conference on Intelligence in Next Generation Networks (ICIN)*, pages 202–209, Piscataway, NJ, 2012. IEEE.

83 Michalis Mountantonakis and Yannis Tzitzikas. Large-scale semantic integration of linked data: A survey. *ACM Comput. Surv.*, 52(5):103:1–103:40, September 2019.

84 Kateryna Neulinger, Anna Hannemann, Ralf Klamma, and Matthias Jarke. A longitudinal study of community-oriented open source software development. In *International Conference on Advanced Information Systems Engineering*, pages 509–523. Springer, 2016.

85 Marcelo Iury S. Oliveira, Glória de Fátima A. Barros Lima, and Bernadette Farias Lóscio. Investigations into data ecosystems: a systematic mapping study. *Knowl. Inf. Syst.*, 61(2):589–630, 2019.

86 Jan Ondrus, Avinash Gannamaneni, and Kalle Lyytinen. The impact of openness on the market potential of multi-sided platforms: A case study of mobile payment platforms. *Journal of Information Technology*, 30(3):260–275, 2015.

87 B Otto, S Lohmann, S Auer, G Brost, J Cirullies, A Eitel, T Ernst, C Haas, M Huber, C Jung, et al. Reference architecture model for the industrial data space. *Fraunhofer-Gesellschaft, Munich*, 2017.

88 Boris Otto. Data ecosystems – conceptual foundations, constituents and recommendations for action.

89 Boris Otto and Matthias Jarke. Designing a multi-sided data platform: findings from the international data spaces case. *Electronic Markets*, 43(1):39, 2019.

90 Margherita Pagani. Digital business strategy and value creation: framing the dynamic cycle of control points. *Mis Quarterly*, pages 617–632, 2013.

91 Geoffrey Parker and Marshall Van Alstyne. Innovation, openness, and platform control. *Management Science*, 64(7):3015–3032, 2017.

92 Thomas F. J.-M. Pasquier and D. Eyers. Information flow audit for transparency and compliance in the handling of personal data. In *2016 IEEE International Conference on Cloud Engineering Workshop (IC2EW)*, pages 112–117, April 2016.

93 Frank Piller. Digitale chancen und bedrohungen – geschäftsmodelle für industrie 4.0. In *VDI Statusreport*. VDI Verlag, 2016.

94 Christoph Pinkel, Carsten Binnig, Ernesto Jiménez-Ruiz, Evgeny Kharlamov, Andriy Nikolov, Andreas Schwarte, Christian Heupel, and Tim Kraska. Incmap: A journey towards ontology-based data integration. In *Datenbanksysteme für Business, Technologie und Web (BTW 2017), 17. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), 6.-10. März 2017, Stuttgart, Germany, Proceedings*, pages 145–164, 2017.

95 Michael E Porter and James E Heppelmann. How smart, connected products are transforming competition. *Harvard business review*, 92(11):64–88, 2014.

96 Christoph Quix, Rihan Hai, and Ivan Vatov. GEMMS: A generic and extensible metadata management system for data lakes. In *28th International Conference on Advanced Information Systems Engineering (CAiSE 2016)*, pages 129–136, 2016.

97 Jean-Charles Rochet and Jean Tirole. Platform competition in two-sided markets. *Journal of the European economic association*, 1(4):990–1029, 2003.

98 Jean-Charles Rochet and Jean Tirole. Two-sided markets: a progress report. *The RAND journal of economics*, 37(3):645–667, 2006.

99 Julian Schütte and Gerd Stefan Brost. A data usage control system using dynamic taint tracking. In *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, pages 909–916, March 2016.

100 Julian Schütte and Gerd Stefan Brost. LUCON: data flow control for message-based iot systems. In *17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications / 12th IEEE International Conference On Big Data Science And Engineering, TrustCom/BigDataSE 2018, New York, NY, USA, August 1-3, 2018*, pages 289–299, 2018.

101 Judith Gebauer Frank Farber Arie Segev. Internet-based electronic markets. *Electronic Markets*, 9(3):138–146, 1999.

102 Srinath Setty. Spartan : Efficient and general-purpose zkSNARKs without trusted setup. *Https://Eprint.Iacr.Org/2019/550*, 2019.

103 K. Su, J. Li, and H. Fu. Smart city and the applications. In *2011 International Conference on Electronics, Communications and Control (ICECC)*, pages 1028–1031, Sep. 2011.

104 B. Tan, S. L. Pan, X. Lu, and L. Huang. The role of is capabilities in the development of multi-sided platforms: the digital ecosystem strategy of alibaba.com. *Journal of the Association for Information systems*, 16(4):248–280, 2015.

105 Felix Ter Chian Tan, Barney Tan, and Shan L. Pan. Developing a leading digital multi-sided platform: Examining it affordances and competitive actions in alibaba.com. *Communications of the Association for Information Systems*, 38:738–760, 2016.

106 Nicolas Tempelmeier, Stefan Dietze, and Elena Demidova. Crosstown traffic - supervised prediction of impact of planned special events on urban traffic. *GeoInformatica. An International Journal on Advances of Computer Science for Geographic Information Systems*, 2019.

107 Nicolas Tempelmeier, Udo Feuerhake, Oskar Wage, and Elena Demidova. St-discovery: Data-driven discovery of structural dependencies in urban road networks. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL 2019, Chicago, IL, USA, November 5-8, 2019*, pages 488–491, 2019.

108 Nicolas Tempelmeier, Yannick Rietz, Iryna Lishchuk, Tina Kruegel, Olaf Mumm, Vanessa Miriam Carlow, Stefan Dietze, and Elena Demidova. Data4urbanmobility: Towards holistic data analytics for mobility applications in urban regions. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 137–145, 2019.

109 David Tilson, Kalle Lyytinen, and Carsten Sørensen. Research commentary—digital infrastructures: The missing is research agenda. *Information systems research*, 21(4):748–759, 2010.

110 Amrit Tiwana, Benn Konsynski, and Ashley A. Bush. Research commentary –platform evolution: Coevolution of platform architecture, governance, and environmental dynamics. *Information Systems Research*, 21(4):675–687, 2010.

111 Maria-Esther Vidal, Kemele M. Endris, Samaneh Jazashoori, Ahmad Sakor, and Ariam Rivas. Transforming heterogeneous data into knowledge for personalized treatments - A use case. *Datenbank-Spektrum*, 19(2):95–106, 2019.

112 Riad S. Wahby, Ioanna Tzialla, Abhi Shelat, Justin Thaler, and Michael Walfish. Doubly-Efficient zkSNARKs Without Trusted Setup. In *Proceedings - IEEE Symposium on Security and Privacy*, 2018.

113 Jonathan Wareham, Paul B Fox, and Josep Lluís Cano Giner. Technology ecosystem governance. *Organization Science*, 25(4):1195–1215, 2014.

114 Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, 1992.

115 Youngjin Yoo, Ola Henfridsson, and Kalle Lyytinen. Research commentary—the new organizing logic of digital innovation: an agenda for information systems research. *Information systems research*, 21(4):724–735, 2010.

Participants

- Cinzia Cappiello
Polytechnic University of Milan, IT
- Ugo de' Liguoro
University of Turin, IT
- Yuri Demchenko
University of Amsterdam, NL
- Elena Demidova
Leibniz Universität Hannover, DE
- Boris Düdder
University of Copenhagen, DK
- Bernadette Farias Lóscio
Federal University of Pernambuco – Recife, BR
- Avigdor Gal
Technion – Israel Institute of Technology – Haifa, IL
- Sandra Geisler
Fraunhofer FIT – Sankt Augustin, DE
- Benjamin Heitmann
Fraunhofer FIT – Aachen, DE & RWTH Aachen, DE
- Fritz Henglein
Univ. of Copenhagen, DK & Deon Digital – Zürich, CH
- Matthias Jarke
RWTH Aachen, DE
- Jan Jürjens
Universität Koblenz-Landau, DE
- Maurizio Lenzerini
Sapienza University of Rome, IT
- Wolfgang Maaß
Universität des Saarlandes – Saarbrücken, DE
- Paolo Missier
Newcastle University, GB
- Boris Otto
Fraunhofer ISST – Dortmund, DE & TU Dortmund, DE
- Elda Paja
IT University of Copenhagen, DK
- Barbara Pernici
Polytechnic University of Milan, IT
- Frank Piller
RWTH Aachen, DE
- Andreas Rausch
TU Clausthal, DE
- Jakob Rehof
TU Dortmund, DE
- Simon Scerri
Fraunhofer IAIS – Sankt Augustin, DE
- Julian Schütte
Fraunhofer AISEC – München, DE
- Egbert Jan Sol
TNO – Eindhoven, NL
- Gerald Spindler
Georg August Universität – Göttingen, DE
- Maria-Esther Vidal
TIB – Hannover, DE

Report from Dagstuhl Seminar 19401

Comparative Theory for Graph Polynomials

Edited by

Jo Ellis-Monaghan[1], Andrew Goodall[2], Iain Moffatt[3], and Kerri Morgan[4]

1 **Saint Michael's College – Colchester, US,** jellismonaghan@gmail.com
2 **Charles University – Prague, CZ,** goodall.aj@gmail.com
3 **Royal Holloway University of London – Egham, GB,** iain.moffatt@rhul.ac.uk
4 **Deakin University – Melbourne, AU,** kerri.morgan@deakin.edu.au

Abstract

This report documents the programme and outcomes of Dagstuhl Seminar 19401 "Comparative Theory for Graph Polynomials".

The study of graph polynomials has become increasingly active, with new applications and new graph polynomials being discovered each year. The genera of graph polynomials are diverse, and their interconnections are rich. Experts in the field are finding that proof techniques and results established in one area can be successfully extended to others. From this a general theory is emerging that encapsulates the deeper interconnections between families of graph polynomials and the various techniques, computational approaches, and methodologies applied to them.

The overarching aim of this Seminar was to exploit commonalities among polynomial invariants of graphs, matroids, and related combinatorial structures. Model-theoretic, computational and other methods were used in order to initiate a comparative theory that collects the current state of knowledge into a more cohesive and powerful framework.

Seminar September 29 – October 4, 2019 – http://www.dagstuhl.de/19401

2012 ACM Subject Classification Mathematics of computing → Graph theory, Mathematics of computing → Graph coloring, Mathematics of computing → Matchings and factors, Mathematics of computing → Graphs and surfaces, Mathematics of computing → Matroids and greedoids, Mathematics of computing → Combinatorial algorithms, Theory of computation → Complexity classes

Keywords and phrases graph polynomials, graph and matroid invariants, Tutte polynomial, topological and algebraic graph theory

Digital Object Identifier 10.4230/DagRep.9.9.135

1 Executive Summary

Jo Ellis-Monaghan (Saint Michael's College – Colchester, US)
Andrew Goodall (Charles University – Prague, CZ)
Iain Moffatt (Royal Holloway University of London, GB)
Kerri Morgan (Deakin University – Melbourne, AU)

This 5-day Seminar built on the previous Dagstuhl Seminar 16241 together with several intervening workshops on graph polynomials, particularly those associated with William Tutte's Centenary, to advance an emerging comparative theory for graph polynomials. Graph polynomials have played a key role in combinatorics and its applications, having

Comparative Theory for Graph Polynomials, *Dagstuhl Reports*, Vol. 9, Issue 9, pp. 135–155
Editors: Jo Ellis-Monaghan, Andrew Goodall, Iain Moffatt, and Kerri Morgan
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

effected breakthroughs in conceptual understanding and brought together different strands of scientific thought. For example, the characteristic and matching polynomials advanced graph-theoretical techniques in chemistry; and the Tutte polynomial married combinatorics and statistical physics, and helped resolve long-standing problems in knot theory. The area of graph polynomials is incredibly active, with new applications and new graph polynomials being discovered each year. However, the resulting plethora of techniques and results urgently requires synthesis. Beyond catalogues and classifications we need a comparative theory and unified approaches to streamline proofs and deepen understanding.

The Seminar provided a space for the cross-fertilization of ideas among researchers in graph theory, algebraic graph theory, topological graph theory, computational complexity, logic and finite model theory, and biocomputing and statistical mechanics applications. There is a long history in this area of results in one field leading to breakthroughs in another when techniques are transferred, and this workshop leveraged that paradigm. More critically, experts in the field have recently begun noticing strong resonances in both results and proof techniques among the various polynomials. The species and genera of graph polynomials are diverse, but there are strong interconnections: in this seminar we worked towards a general theory that brings them together under one family. The process of developing such a theory of graph polynomials exposes deeper connections, giving great impetus to both theory and applications. This has immense and exciting potential for all those fields of science where combinatorial information needs to be extracted and interpreted.

The seminar was roughly organized according to the following themes:

- **Unification:** General frameworks for graph polynomials including meta-problems, K-theory, Second Order Logic, and Hopf algebras.
- **Generalizations:** Polynomial invariants for graphs with added structure (e.g. digraphs, ribbon graphs) or more general "underlying" combinatorial structures (e.g. matroids, Δ-matroids).
- **Distinction:** Distinguishing power of graph invariants (equivalence and uniqueness up to isomorphism with respect to a given graph polynomial, interrelations among graph polynomials, properties of graph polynomials).
- **Applications:** Applications of graph polynomials in other disciplines (e.g. self-assembly, sequencing, quantum walks, statistical mechanics, knot theory, quantum Ising model).
- **Conjectures:** Breakthrough conjectures (outstanding open problems whose resolution would have a broad impact on the understanding of graph polynomials).
- **Complexity:** Computational complexity and computational methods.

2 Table of Contents

3 Structure of the Workshop

Beyond simple information exchange, the goal of this workshop was to make concrete progress towards identified problems that would move the field forward. Thus, we adopted the following ‘working group’ format for the week. This format optimized collaboration time, but also included a number of plenary talks. We particularly note that the social spaces were fully crowded every night of the Seminar, with nearly all the participants reconvening after dinner to continue a freely flowing exchange of ideas about the themes of the Seminar.

- **Monday:**
  - The workshop began with an opening plenary talk, touching on the foundations of the field with emphasis on untapped open areas from seminal papers and long-standing conjectures.
  - After the plenary talk, the organizers gave an overview of open problems collected from other events, organized by the broad themes of the workshop: unification, generalizations, distinction, complexity, applications and conjectures.
  - This was followed by 5-minute presentations of new open problems, again grouped according to the workshop themes.
  - Presentations of open problems were completed shortly after lunch, after which participants self-selected into the six working groups following the broad themes of the workshop. The working groups adjourned to their breakout rooms to begin work on the identified open problems, and in many cases new problems were generated by further discussion within the working groups.
- **Tuesday:**
  - The day began with short invited talks by early career researchers.
  - The remainder of the day was dedicated to intensive collaborative work in working groups. This extended period of concentrated time led to meaningful progress in collaborative efforts.
- **Wednesday:**
  - The day began with short interim progress reports from the working groups and any subgroups that formed within them. Interim reports served to establish working notes that increase the likelihood of research continuance after workshop. At this juncture, there was also some movement of participants among groups, and one or two subgroups moved on to other projects.
  - The rest of the morning was spent collaborating in the working groups.
  - In the afternoon about half the participants continued their investigations while the other half went on an excursion to Völklinger Hütte.
- **Thursday:**
  - The entire day was devoted to working groups. Some groups reported that this concentrated, uninterrupted time was a significant factor in resolving their selected problems and framing out papers for publication.
  - On Thursday night at dinner we solicited written responses from each table as to where the field should go from here, what overarching questions would drive it there, and how we might continue the efforts of this workshop.
- **Friday:**
  - In the first morning session, the working groups consolidated their working notes, prepared their final reports, and made concrete plans for continuing their efforts to bring their work at the workshop to fulfilment.

- After the coffee break, each working group and subgroup presented their outcomes to the whole group, and shared plans for continuing their work. These plans ranged from simply completing papers now well-underway to planning follow up workshops, and most included agreeing mechanisms for continuing collaboration (Dropbox, Skype, etc.).
- After the reports, we requested and received verbal feedback from the participants on what worked well for them at the workshop and what might improve it. The workshop concluded with a whole-group discussion about the next steps that we should take as a community to best enhance the subject.

4 Overview of Plenary Talks

4.1 Tutte–Whitney polynomials: some history and problems

Graham Farr (Monash University – Clayton, AU)

Tutte-Whitney polynomials are algebraic data structures that contain a lot of important information about graphs. Their evaluations and specialisations include, for example, the chromatic, flow and reliability polynomials of a graph, the number of spanning trees, numbers of acyclic and totally cyclic orientations, the Ising and Potts model partition functions of statistical physics, the weight enumerator of a linear code, and the Jones polynomial of an alternating link.

We describe the history of Tutte-Whitney polynomials, especially the contributions of Hassler Whitney and W. T. (Bill) Tutte, and with some emphasis on Tutte's 1947 paper, 'A ring in graph theory', parts of which are still not well known. We find unexpectedly early occurrences in the literature of several important results, concepts and questions, and take several opportunities to comment on computational aspects of the theory.

This history sets the scene for a number of open problems and proposed future research directions. These include: a question of Whitney about a possible strengthening of the Four-Colour Theorem; some questions about equivalence of graphs with respect to particular graph polynomials (e.g., chromatic equivalence), in particular some questions about certificates of chromatic equivalence; some questions about factorisation of graph polynomials (e.g., chromatic factorisation) and certificates of chromatic factorisation; trying to link these theories of certificates to work on rewriting systems, word problems, and computational algebra.

4.2 Transversal polynomial of covers of graphs

Krystal Guo (UL – Brussels, BE)

Joint work of Chris Godsil, Krystal Guo, Gordon Royle

We study a polynomial with connections to correspondence colouring, a recent generalization of list colouring, and the Unique Games Conjecture. Given a graph G and an assignment of elements of the symmetric group S_r to the edge of G, we define a cover graph: there are sets of r vertices corresponding to each vertex of G, called fibres, and for each edge uv, we add a perfect matching between the fibres corresponding to u and v. A transversal subgraph of the cover is an induced subgraph which has exactly one vertex in each fibre. In this setting, we can associate correspondence colourings with transversal cocliques and unique label covers with transversal copies of G.

We define a polynomial which enumerates the transversal subgraphs of G with k edges. We show that this polynomial satisfies a contraction deletion formula and use this to study the evaluation of this polynomial at $-r+1$.

4.3 Interpretations of the Tutte Polynomials of Regular Matroids

Martin Kochol (Bratislava, SK)

A regular chain group N is the set of integral vectors orthogonal with rows of a matrix representing a regular matroid M, i.e., a totaly unimodular matrix. N corresponds to the set of flows and tensions if M is a graphic and cographic matroid, respectively. We evaluate the Tutte polynomial of M as number of pairs of specified elements of N.

4.4 Matching polynomials

Bodo Lass (University Claude Bernard – Lyon, FR)

We use the commutative algebra of set functions to prove results about matching polynomials.

4.5 Complexity of evaluation of graph polynomials

Johann A. Makowsky (Technion – Haifa, IL)

Joint work of Andrew Goodall, Miki Hermann, Tomer Kotek, Johann A. Makowsky, Steven Noble

We define the complexity spectrum of graph polynomials and collect results from the literature, including our own, which completely describe the complexity spectrum of classical graph polynomials, such as the Tutte polynomial and its many variations and instantiations, the matching polynomial, the interlace polynomial and the cover polynomial. We then concentrate on univariate graph polynomials, especially Harary polynomials which count the number of various colorings with k colors. Besides exhibiting possible behaviour of complexity spectra we also formulate open problems, solutions of which we think should be in the range of our current capabilities.

4.6 Acyclic orientation polynomials and the sink theorem for the chromatic symmetric function

Jaeseong Oh (Seoul National University, KR)

Joint work of Byung-Hak Hwang, Woo-Seok Jung, Kang-Ju Lee, Jaeseong Oh, Sang-Hoon Yu

We define the acyclic orientation polynomial of a graph to be the generating function for the sinks of its acyclic orientations, which is a refinement of the number of acyclic orientations. We show that our acyclic orientation polynomial satisfies a deletion-contraction recurrence with a change of variables. As the main application, we provide a new proof for Stanley's sink theorem for the chromatic symmetric function X_G, which gives a relation between the number of acyclic orientations of G with a fixed number of sinks and the coefficients in the expansion of X_G with respect to elementary symmetric functions.

5 A Selection of Open Problems and Questions Presented at the Seminar

Some three dozen problems were presented at the start of the Seminar, with more emerging during the breakout sessions. Some were solved during during the Seminar with papers drafted, while significant progress was made towards others. This section contains a small selection of the problems presented.

5.1 Graph polynomials through combinatorial Hopf algebra

Nantel Bergeron (York University – Toronto, CA)

We can construct the Stanley Chromatic Symmetric function using combinatorial Hopf algebras. In fact, for a graph G on vertex set $[n] = \{1, 2, \dots n\}$, we get a quasisymmetric function (that is symmetric since the Hopf algebra is cocommutative):

$$\Psi(G) = \sum_{A \models [n]} \zeta(G|_{A_1}) \cdots \zeta(G|_{A_k}) M_{\alpha(A)},$$

where the sum is over all set composition $A = (A_1, A_2, \dots, A_k)$ of the set $[n]$, ζ is 1 if the input graph has no edges and 0 otherwise, and $M_{\alpha(A)}$ is the monomial quasisymmetric function indexed by the integer composition $\alpha(A) = (|A_1|, |A_2|, \dots, |A_k|)$. Stanley and Stembridge have conjectured that $\Psi(G)$ is e-positive if G is the incomparable graph of a $3+1$ avoiding poset, and that it is Schur positive if G is claw free. We have recently shown that $\Psi(G)$ is always positively h-alternating (the coefficient of $(-1)^{n-\ell(\lambda)} h_\lambda$ is always positive). This last result is a refinement of the alternation of the chromatic polynomial.

Question 1

We can choose different characters (analogue of *zeta* above) and produce different invariants for graphs. Can we find analogue of the Stanley and Stembridge conjecture?

Question 2

We can do this in larger Hopf algebras (For examples hypergraphs). For an arbitrary hypergraph H, it is not true that $\Psi(H)$ is positively h-alternating. Can we characterize families of hypergraph with the property that $\Psi(H)$ is positively h-alternating. Maybe something like for any edge $E \in H$ and any contraction H/S (with respect to a subset of edges S), the cardinality of E/S is even or 1? What can we say for matroids?

5.2 Recognition

Anna De Mier (UPC – Barcelona, ES)

A short statement of the problem is “Does the Tutte polynomial recognize (matroid) parallel connection?”

This question is motivated by the fact that to my knowledge we still do not know whether the Tutte polynomial can tell apart 3-connected graphs (that is, whether any graph Tutte-equivalent to a 3-connected one is also 3-connected). It is well known, though, that a 3-connected matroid can have the same Tutte polynomial as a non-3-connected one. By going through examples of Tutte-equivalent matroids I have not found any pair where one is 3-connected and the other one is a parallel connection (in the known examples mentioned above, then non-3-connected one is a 2-sum, but not a parallel connection). So this raises the question of whether the Tutte polynomial “sees” parallel connections, somehow in the same way that it does “see” direct sums. That is, we would like to know if there is a pair of Tutte-equivalent matroids M and N such that M can be written as a parallel connection but N cannot.

I would be also interested in the answer of the previous question for the G-invariant, or for any invariant stronger than the Tutte polynomial.

5.3 Some overarching ‘meta-problems’ for unification

Jo Ellis-Monaghan (Saint Michael’s College – Colchester, US)

What attributes might a unified theory for graph polynomials have?

- Some desirable properties:
  - a common framework encompassing existing graph polynomials,
  - underlying (algebraic?) structure that reveals their relations,
  - a hierarchy or partial ordering with respect to the distinguishing power of the various polynomials,
  - common tools, theorems, or conditions for establishing common properties (recursion, generating functions, universality, etc.), and
  - generic computational complexity results.
- Some possible directions:

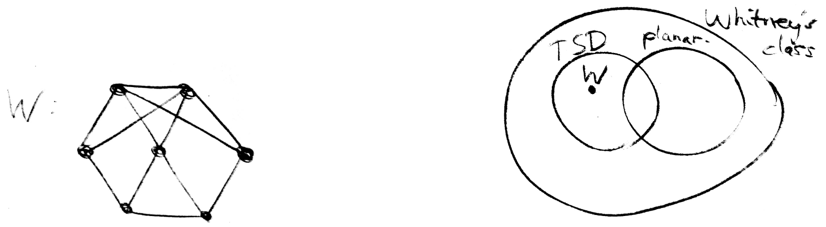


Figure 1 Whitney's example of a non-planar Tutte self-dual graph.

- Characterize graph reductions that give well-defined linear recurrence relations and hence lead to graph polynomials.
- What are the possible mechanisms for comparing graph polynomials? (e.g. via SOL or kernel containments). What partial orders for a class of graph polynomials arise from these? What can we learn from the structure of such a poset?
- A graph polynomial creates equivalence classes of graphs for which it returns the same polynomial. When is it possible to decide for a given equivalence relation of graphs whether there is a graph polynomial that realises it?
- Graph polynomials build bridges between algebraic/enumerative properties and structural/combinatorial properties. Can we make these bridges precise in the field?

5.4 Long-standing problems

Graham Farr (Monash University – Clayton, AU)

An old question of Hassler Whitney (1932) about an elegant extension of the Four Colour Theorem using duality and Tutte–Whitney polynomials.

Let G be a graph such that there exists a graph H with the property that $T(G^*;x,y) = T(H;x,y)$. In other words, G^* is Tutte-equivalent to a graph, even though it might not be a graph itself. Here, G^* is the matroid dual of G, so it is a cographic matroid, but is only graphic if G is planar.

Question: Are such graphs G 4-colourable?

Whitney observed that this was a strengthening of the Four Colour Problem.

- If G is planar then G^* is also a planar graph and we may take $H = G^*$. In this case, the answer to the question is Yes: it's just the Four Colour Theorem.
- A graph G is *Tutte self-dual* (TSD) if $T(G^*;x,y) = T(G;x,y)$, in which case we may take $H = G$. Whitney gave an example of a non-planar Tutte self-dual graph: see W in Figure 1. This graph is 4-colourable.

5.5 Some problems coming from algebra

Alex Fink (Queen Mary University of London, GB)

The one I have the strongest philosophical commitment to:

Problem 1

Consider the Tutte polynomial, say. It's known exactly which of its evaluations are efficiently computable. Which *partial derivatives* of each evaluation are efficiently computable?

Problem 2

Once we know this, we know the largest quotient of $Z[x, y]$ in which the image of Tutte is efficiently computable. Now look at the structure of that quotient ring and see if this evaluation can be characterised as the universal invariant satisfying meaningful recurrences vel sim. (The philosophy is stop thinking only of maps to polynomial rings: there's lots more rings out there.)

The obvious one: Milk more invariants of interest out of the machinery of arXiv:1711.09028 and arXiv:1508.00814.

Problem 3

One from last time: Construct a meaningful trivariate polynomial of a three-coloured triangulation of the sphere which reflects the triality of Kalman and Posnikovos univariate polynomial under permuting colours, and specialises to my and Amanda Cameron's bivariate polynomial.

Problem from last time

Let G be a connected bipartite graph and $V_\mathrm{e} \amalg V_\mathrm{v}$ its vertex set. A *hypertree* for G is the degree sequence in $Z^{|V_\mathrm{e}|}$ of some spanning tree of G (these form a *hypergraphic polymatroid*). Define the bivariate polynomial $Q(G; t, u)$ so that, when t and u are naturals,

$$\begin{aligned} Q(G; t, u) = \#\{p \in Z^{|V_\mathrm{e}|} : p = a + b + c, \quad & a \text{ is a hypertree of } G, \\ b_i \in Z_{\leq 0}, \ \sum_i b_i = -t, \text{ and } \quad & c_i \in Z_{\geq 0}, \ \sum_i c_i = u\}. \end{aligned}$$

Ehrhart theory guarantees the existence of this polynomial. When all vertices in V_e are bipartite, then G is the barycentric subdivision of a graph H; in this case, hypertrees for G are in bijection with spanning trees for H, and $Q(G; t, u)$ contains the same information as $T(H; x, y)$. To wit, with Amanda Cameron we've shown that

$$\sum_{t,u \geq 0} Q(G; t, u)\alpha^t \beta^u = \frac{T\left(H; \dfrac{1-\alpha\beta}{1-\beta}, \dfrac{1-\alpha\beta}{1-\alpha}\right)}{(1-\alpha)^{|V(H)-1|}(1-\beta)^{|E(H)-V(H)+1|}(1-\alpha\beta)}.$$

Now let Δ be a three-coloured triangulation of the sphere. Then there are six ways to delete one colour class from Δ, leaving a bipartite graph G, and label the other two colour classes V_{e} and V_{v}. If V_{e} is colour i, V_{v} is colour j, and the deleted colour is k, let $Q_{ijk}(\Delta; t, u) = Q(G; t, u)$. These are interrelated. Firstly,

$$Q_{ijk}(\Delta; t, u) = Q_{ikj}(\Delta; u, t).$$

In the case where all vertices of colour i have degree 4, this is plane graph duality (in general, it's a polymatroid duality). Secondly, Kálmán and Postnikov [1] have shown that

$$Q_{ijk}(\Delta; t, 0) = Q_{jik}(\Delta; t, 0).$$

This is all compatible with the existence of a trivariate polynomial $\widehat{Q}(\Delta; x_i, x_j, x_k)$ such that

$$\widehat{Q}(\Delta; 0, x_j, x_k) = Q_{ijk}(\Delta; x_k, x_j)$$

and such that permuting the colour classes of Δ permutes the variables of $\widehat{Q}(\Delta)$ in the corresponding way.
Problem: Construct a nice such $\widehat{Q}(\Delta)$.

References

1 T. Kálmán and A.Postnikov, Root polytopes, Tutte polynomials, and a duality theorem for bipartite graphs, arXiv:1602.04449.

5.6 Filtrations and decompositions

Emeric Gioan (University of Montpellier & CNRS, FR)

Let $G = (V, E)$ be a graph and suppose that E is linearly ordered. A *connected filtration* of G consists of a sequence of sets

$$\emptyset = F'_j \subset \ldots \subset F'_0 = F_c = F_0 \subset \ldots \subset F_i = E$$

such that:

1. the sequence $min(F_k \setminus F_{k-1})$, $1 \leq k \leq i$, is increasing with k;
2. the sequence $min(F'_{k-1} \setminus F'_k)$, $1 \leq k \leq j$, is increasing with k;
3. for every $1 \leq k \leq i$, the minor $G(F_k)/F_{k-1}$ is either a single isthmus (if it has one edge) or is loopless and 2-connected (otherwise);
4. for every $1 \leq k \leq j$, the minor $G(F'_{k-1})/F'_k$ is either a single loop (if it has one edge) or is loopless and 2-connected (otherwise).

When the two last conditions are omitted, we define a *filtration* of G.

Consider the following result.
Theorem [G. & Las Vergnas, 2002, 2019]
With $\beta^*(G) = \beta(G)$ if $|E| > 1$, $\beta^*(\text{isthmus}) = 0$, $\beta^*(\text{loop}) = 1$, we have

$$t(G; x, y) = \sum_{\substack{\text{(connected)}\\ \text{filtrations}}} \Big(\prod_{1 \leq k \leq i} \beta\big(G(F_k)/F_{k-1}\big)\Big)\Big(\prod_{1 \leq k \leq j} \beta^*\big(G(F'_{k-1})/F'_k\big)\Big) x^i y^j.$$

Note that this result "contains" the spanning tree activity formula, the orientation activity formula, the duality formula, and the convolution formula. It is also related to a unique (connected) filtration canonically decomposing a spanning tree or a directed graph. It is probably related to Hopf algebras (Krajewski, Moffatt, Tanasa, Dupont, Fink, Moci) and to Bergman/conormal fans in matroid geometry (Ardila et al., Rincon et al.), and it can be seen in terms of the algebra of set functions (Lass). Actually, this result is more generally available in matroids as well.

We ask the following questions. What are the structural or computational uses of the underlying decompositions? Are there related meta-results without fixing a linear ordering? Are there similar approaches and results for other graph polynomials?

5.7 Psicle polynomials

Krystal Guo (UL – Brussels, BE)

We define a polynomial which interpolates between the characteristic polynomial of a graph and the matching polynomial of a graph. The polynomial takes, as input, a set of cycles of the graph, which can be chosen to be the set of contractible cycles of a given embedding. There polynomials gives a natural walk-generating function for certain types of walks in the graph. There are interesting questions about when this polynomial has real roots, or if the roots lie in a restrict section of the plane. More interestingly, we can ask whether or not this property is related to the orientability of the surface.

A graph is *sesquivalent* if all of its components are 1- or 2-regular; that is to say, a graph is sesquivalent if it is a disjoint union of cycles and copies of K_2. Let X be a graph on n vertices and let $\Gamma_r(X)$ be the set of all sesquivalent subgraphs of X on r vertices, for a fixed $0 \leq r \leq n$. For a sesquilinear graph Y, we define $\overline{Y}$ to be the number of components of Y and $\langle Y \rangle$ to be the number of 2-regular components of Y. By expanding the determinant of $tI - A(X)$, we see that the coefficient of x^{n-r} in the characteristic polynomial $\phi(X,t)$ of X is as follows:

$$\sum_{Y \in \Gamma_r(X)} (-1)^{\overline{Y}} 2^{\langle Y \rangle}.$$

We explore a polynomial of X with a similar expansion, in terms of sesquivalent subgraphs of X, where we only sum over all sesquivalent subgraphs whose cycles are in a set of allowable cycles.

Let C be a subset of all cycles of X. For a fixed $0 \leq r \leq n$, let $\Gamma_r(X, C)$ be the set of all sesquivalent subgraphs Y of X on r vertices, such that each 2-regular component of Y is in C. We define the *psicle polynomial* of X with respect to C to be the polynomial $\psi(X, C, t)$ such that the coefficient of x^{n-r} is as follows:

$$\sum_{Y \in \Gamma_r(X,C)} (-1)^{\overline{Y}} 2^{\langle Y \rangle}.$$

Observe that if we choose C to be the set of all cycles of X, then $\psi(X, C, r) = \phi(X, t)$. If $C = \emptyset$, then $\psi(X, C, r) = \mu(X, t)$, where $\mu(X, t)$ denotes the matching polynomial of X.

Given a cellular embedding of X on a surface Σ, a sensible choice for a subset of cycles C_c would be the set of contractible cycles of X. We define the *psicle polynomial of* X *with respect to an embedding* Π to be $\psi(X, C_c, t)$, where C_c is the set of contractible cycles of X with respect to Π. We will abuse notation and denote this as $\psi(X, \Pi, t)$.

We would like to investigate the location of the roots of $\psi(X, \Pi, t)$ in a complex plane.

1. For a given graph X, what are some properties of embeddings Π such that $\psi(X, \Pi, t)$ has real roots? Does orientability matter?
2. Is it true that every complex root of $\psi(X, \Pi, t)$ has negative real part? (Or a similar statement?)

5.8 Some problems coming from knot theory

Louis H. Kauffman (Novosibirsk State University, RU)

Loops (even with virtual crossings) receive the value d (the third variable). S. Baldridge, L. Kauffman and W. Rushworth are studying a mapping from virtual 3-valent graphs G with perfect matching M to virtual link diagrams. This mapping takes a natural perfect matching polynomial for graphs (generalizing the Penrose evaluation) to the three variable bracket, and hence to the Jones polynomial by specialization.

This is a new relationship between graph polynomials and link invariants, and it goes both ways, allowing essentially all invariants of virtual knots to be viewed as invariants of perfect matching graphs up to a pullback to graphs of knot theoretic isotopy. More generally, I am working on the interface between virtual knot theory, graph theory and statistical mechanics and quantum theory via the above construction and via generalizations of the medial constructions that have a long history in this subject, via surfaces and ribbon graphs and via the category of Morse diagrams, quantum link invariants and link homology. The purpose of this five-minute talk is to give a very quick introduction to the new mapping from perfect matching graphs to virtual links and to indicate the many problems and constructions that this entails.

References

1 S. Baldridge, L.H. Kauffman, W. Rushworth, Graphenes and Virtual Knots and Links, (in preparation).

2 S.Baldridge.A cohomology theory for planar trivalent graphs with perfect matchings. 2018. arXiv. org/abs/1810.07302.

3 S. Baldridge, A. Lowrance, and B. McCarty. The 2-factor polynomial detects even perfect matchings. 2018. arXiv:1812.10346.

4 J. A. Ellis-Monaghan, L. H. Kauffman, I. Moffatt, Edge colourings and topological graph polynomials. Australas. J. Combin. 72 (2018), 290-305.

5 L.H. Kauffman, State Models and the Jones Polynomial, Topology 26 (1987), 395-407.

6 L. H. Kauffman, Virtual Knot Theory. European J. Combin., 20(7):663-691, 1999.

7 L. H. Kauffman, Knots and Physics World Scientific Pub. Co., Singapore (1991, 1994, 2012).

8 Dye, Heather A.; Kaestner, Aaron; Kauffman, Louis H. Khovanov homology, Lee homology and a Rasmussen invariant for virtual knots. J. Knot Theory Ramifications 26 (2017), no. 3, 1741001, 57 pp.

9 L. H. Kauffman, A self-linking invariant of virtual knots. Fund. Math. 184 (2004), 135-158, math.GT/0405049.

10 L. H. Kauffman, Knot diagrammatics. Handbook of Knot Theory, edited by Menasco and Thistlethwaite, 233?318, Elsevier B. V., Amsterdam, 2005. math.GN/0410329.

11 L. H. Kauffman, Introduction to virtual knot theory. J. Knot Theory Ramifications, 21(13), 2012.

12 L. H. Kauffman, Map coloring and the vector cross product J Comb Theo B vol 48, no 2, April (1990) 145-154.

13 L. H. Kauffman, Reformulating the map color theorem. Discrete Math. 302 (2005), no. 1-3, 145-172.

14 L. H. Kauffman, Map Coloring, q-Deformed Spin Networks, and Turaev–Viro Invariants for 3-Manifolds. Intl. J. Mod. Phys. B, Vol. 6, Nos. 11, 12 (1992), p. 1765-1794.

15 L. H. Kauffman, A state calculus for graph coloring, Illinois Journal of Mathematics, Volume 60, Number 1, Spring 2016, Pages 251271 S 0019-2082 (Special Issue of the Illinois J. Math. celebrating Wolfgang Haken).

16 V.O.Manturov, Khovanov homology for virtual links with arbitrary coeffcients. J. Knot Theory Ramifications, 16(03):343377, 2007.

17 R. Penrose, Applications of negative dimensional tensors, in Combinatorial Mathematics and Its Applications edited by D. J. A. Welsh, Acad. Press (1971).

18 E. Witten, Quantum Field Theory and the Jones Polynomial. Comm. in Math. Phys. Vol. 121 (1989), 351-399.

5.9 Graph polynomials and (n, r)-matroids

Joseph Kung (University of North Texas – Denton, US)

The Tutte polynomial of a rank r matroid on n elements has degree r as a polynomial in x and degree $n - r$ as a polynomial in y. The Tutte polynomials of such matroids span a subspace of $\mathbb{C}[x, y]$ and an upper bound for

$$\dim\langle\, T(M; x, y) : M \text{ an } (n, r)\text{-matroid}\,\rangle$$

is $(r + 1)(n - r + 1)$.

The dimension is in fact equal to $r(n - r) + 1$.

Problem: Answer the same question for other graph polynomials.
The dimension of the subspace spanned by the graph polynomial for graphs of given order and size serves as a measure of information contained in a graph polynomial: how useful is this way of measuring the combinatorial information contained in a given polynomial graph invariant?

5.10 Graph polynomial problems arising from Vassiliev and quantum knot invariants

Sergei Lando (NRU Higher School of Economics – Moscow, RU)

Problem 1: The existence of the Lie algebra polynomial graph invariant

A weight system is a function on chord diagrams (circles with a tuple of chords without common ends) satisfying certain conditions called 4-term relations. In particular, Bar-Natan and Kontsevich associated a weight system to a semi-simple Lie algebra endowed with an invariant nondegenerate scalar product. To a chord diagram, its intersection graph can be associated: the vertices of the graph are the chords of the diagram, and two vertices are connected by an edge if and only if the corresponding chords intersect one another. Not every graph arises as the intersection graph of a chord diagram, and certain graphs are intersection graphs of several chord diagrams.

In [1], it is shown that for the Lie algebra sl_2 the value of the corresponding weight system depends on the intersection graph of the chord diagram rather than on the diagram itself. As a result, we obtain a partially defined graph invariant taking values in the ring of polynomials in a single variable (which is the Casimir element of the Lie algebra sl_2).

Question: is there is an extension of this partially defined polynomial graph invariant to a completely defined polynomial graph invariant satisfying 4-term relations for graphs?

The 4-term relations for graphs were introduced in [3]. E. Krasilnikov constructed an extension of this graph invariant to all graphs with up to 8 vertices and proved uniqueness of the extension.

Problem 2: The value of the Lie algebra sl_2 weight system on complete graphs

The graph invariant from the previous problem is well-defined for complete graphs since they are intersection graphs. However, known methods of computation of its value on a complete graph with a given number of vertices, are laborious, and I know the answer only for complete graphs with up to 14 vertices. These results allow me to formulate a conjecture representing the generating function for these polynomials as a continuous fraction.

Problem: prove the conjectural formula for the values of the sl_2*-invariant on complete graphs.*

Problem 3: Umbral invariants for binary delta-matroids

An *umbral polynomial graph invariant* is a graded Hopf algebra homomorphism from the Hopf algebra of graphs to the Hopf algebra of polynomials in infinitely many variables. An example is given by the Stanley symmetrized chromatic polynomial. It is proved in [2] that the mean value of an umbral polynomial graph invariant is, essentially, a linear combination of one-part Schur polynomials. However, a similar statement is not valid for the Hopf algebra of binary delta-matroids.

Question: what should be a correct replacement of the statement for graphs for binary delta-matroids?

References

1 S. Chmutov, S.Lando, Mutant knots and intersection graphs, Algebr. Geom. Topol. 7 (2007), 1579–1598.

2 S. Chmutov, M. Kazarian, S.Lando, Polynomial graph invariants and the KP hierarchy, arXiv:1803.09800.

3 S. Lando, On a Hopf algebra in graph theory, J. Combin. Theory Ser. B 80 (2000), 104–121.

5.11 Matchings and monotonicity

Bodo Lass (University Claude Bernard – Lyon, FR)

Let us consider bipartite graphs $G = (X, Y; E)$, where X is a fixed set of vertices, $|X| = n$, and Y is a variable set of vertices. For every subset X' of X, let us denote by $d(X')$ the number of neighbors (in Y) of X'. Hall's classical theorem on matchings affirms that there is an injective function $f : X \to Y$ such that $\{x, f(x)\}$ is an edge of G for every $x \in X$ if and only if $d(X') \geq |X'|$ for every nonempty subset X' of X.

But one may also want to count the number $m(X)$ of such injective functions (matchings covering X). I have shown that the $2^n - -1$ numbers $d(X')$ (X' nonempty subset of X) determine G up to isomorphism and allow to calculate $m(X)$ in the following way:

We must sum over all partitions of X into nonempty blocks X', where each block X' contributes the factor

$$(|X'| - 1)![d(X') - (n - 1)].$$

For example, if $X = \{1, 2, 3\}$, then

$$m = (d_1 - 2)(d_2 - 2)(d_3 - 2) + (d_{12} - 2)(d_3 - 2) + (d_{13} - 2)(d_2 - 2) + (d_{23} - 2)(d_1 - 2) + 2(d_{123} - 2).$$

This formula shows that m, in general, depends monotonically on the set function d. Because of the substraction of $(n - 1)$, however, this formula does not show monotonicity in all cases. In particular, it does not prove the most extreme case : Hall's theorem. Eberhard Triesch conjectured that the monotonicity is always true, and it would be nice to find a proof or counterexample. More details (in latex and pdf) can be found on http://math.univ-lyon1.fr/~lass/articles/pub17triesch.html.

5.12 Complexity of the evaluation of graph polynomials

Johann A. Makowsky (Technion – Haifa, IL)

Let C be a set of clauses and (V, C, R) a directed graph in which arcs (R) join variables (V) to clauses (C) with direction according to whether the variable is negated or not in the clause. (See Figure 2.)

For $A \subseteq V$, define $\mathrm{SAT}(A) = \{c \in C : A \text{ satisfies } c\}$ and the SAT-polynomial in indeterminate X by

$$\sum_{A \subseteq V} \prod_{c \in \mathrm{SAT}(A)} X = \sum_{A \subseteq V} X^{|\mathrm{SAT}(A)|}.$$

Is this polynomial useful for studying the satisfiability problem SAT?

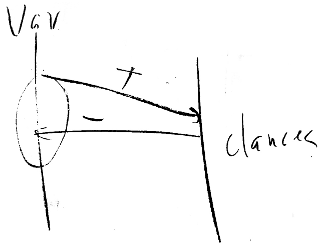


Figure 2 Assignments of values (V) to variables in clauses (C), negative arcs for negated variables.

5.13 Zero-free regions for the chromatic polynomial

Guus Regts (University of Amsterdam, NL)

Determine regions in the complex plane on which the chromatic polynomial does not vanish for interesting classes of (and possibly all) bounded degree graphs. In particular improve on the results from [1]

References

1 R. Fernández and A. Procacci, Regions without complex zeros for chromatic polynomials on graphs with bounded degree, Combin. Probab. Comput. 17 (2008), no. 2, 225–238.

5.14 Distinguishing power: Krushkal vs 4-variable from surface Tutte polynomial

Lluís Vena Cros (Charles University – Prague, CZ)

Goodall, Krajewski, Regts and Vena, and Goodall, Litjens, Regts and Vena recently introduced a polynomial for maps on orientable and non-orientable surfaces. These two polynomials can be seen as a common generalization of the flow polynomial (which counts nowhere-identity flows of a map in which edges are given values in a not necessarily abelian finite group G, and the Kirchhoff law equations are determined by the cyclic orientation of edges around each vertex of the embedding) and the local tension polynomial (which count nowhere-identity flows of the surface-dual map).

The polynomial for maps on orientable surfaces is defined by

$$\mathcal{T}(M;\mathbf{x},\mathbf{y}) = \sum_{A\subseteq E} x^{n^*(M/A)} y^{n(M\backslash A^c)} \prod_{\substack{\text{conn. cpts } M_i\\ \text{of } M/A}} x_{g(M_i)} \prod_{\substack{\text{conn. cpts } M_j\\ \text{of } M\backslash A^c}} y_{g(M_j)}, \qquad (1)$$

where M is the given map on an orientable surface (a graph with a cyclic ordering of edges, actually half-edges, around each vertex), $M\backslash A^c$ is the map where the edges of the complement of A have been deleted (the relative cyclic order of the other half-edges is preserved), and M/A is the map where the edges in A have been map-contracted (map-contraction is the surface-dual operation to edge deletion), and g is the orientable genus of the map (the number of handles attached to a sphere, extended additively over connected components for non-connected maps), $n(M) = e(M) - v(M) + k(M)$ is the nullity of the map (the number of edges minus the number of vertices plus the number of connected components), and $n^*(M) = e(M) - f(M) + k(M)$ is the dual nullity ($f(M)$ is the number of faces of M).

The polynomial is defined more generally for maps M embedded in (not necessarily orientable) surfaces by

$$\mathcal{T}(M;\mathbf{x},\mathbf{y}) = \sum_{A\subseteq E} x^{n^*(M/A)} y^{n(M\backslash A^c)} \prod_{\substack{\text{conn. cpts } M_i\\ \text{of } M/A}} x_{\bar{g}(M_i)} \prod_{\substack{\text{conn. cpts } M_j\\ \text{of } M\backslash A^c}} y_{\bar{g}(M_j)}, \tag{2}$$

where $\bar{g}(M)$ is the "signed genus", equal to the genus when M is orientable, and equal to the negative of Euler's demigenus when M is non-orientable (equal to the number of cross-caps attached to a sphere).

Another polynomial defined for maps is the Krushkal polynomial (defined in a paper from 2011), which has four variables and that can be defined as follows:

$$\mathcal{K}(M;X,Y,C,D) = \sum_{A\subseteq E} X^{k(M\backslash A^c)-k(M)} Y^{k(M/A)-k(M)} C^{s(M\backslash A^c)/2} D^{s(M/A)/2}, \tag{3}$$

where $s(M) = 2k(M) - \chi(M)$ is the Euler genus, with $\chi(M)$ the Euler characteristic of M.

It can be checked that

$$\mathcal{K}(M;X,Y,C,D) = X^{-k(M)} Y^{-k(M)} \mathcal{T}(M;\mathbf{x},\mathbf{y})$$

with $x = 1$, $y = 1$, $x_{\bar{g}} = YD^{\frac{\bar{g}+3|\bar{g}|}{4}}$ and $x_{\bar{g}} = XC^{\frac{\bar{g}+3|\bar{g}|}{4}}$ for $\bar{g} \in \mathbb{Z}$.

There are two maps M_1 and M_2 for which (1) induces two different polynomials, yet they are the same polynomial for (3). Hence, (1) induces a strictly finer partition on the set of maps.

A 4-variable polynomial specialization of (1) and (2) is given by

$$Q(M,x,y,a,b) = \sum_{A\subseteq E} x^{n^*(M/A)} y^{n(M\backslash A^c)} a^{\bar{g}(M/A)} b^{\bar{g}(M\backslash A^c)}.$$

upon taking $x_{\bar{g}} = a^{\bar{g}}$ and $y_{\bar{g}} = b^{\bar{g}}$ for each $\bar{g} \in \mathbb{Z}$.

The question is: Does the polynomial Q have the same, more, less, or different distinguishing power than $\mathcal{K}$?

6 Heidelberg Laureate Forum Participant Talk

Animesh Chaturvedi (PhD student, IIT Indore) who was attending the Heidelberg Laureate Forum (2019) was invited by Schloss Dagstuhl – Leibniz Center to attend the seminar, and presented a short talk.

6.1 System evolution analytics: data mining and learning of evolving graphs representing evolving complex system

Animesh Chaturvedi (Indian Institute of Technology – Indore, IN)

Usually, real-world evolving systems have many entities (or components), which evolves over time. A technique can be used to analyze the evolving system; the technique is named as System Evolution Analytics. Such techniques can be applied on an evolving system represented as a set of temporal networks. These techniques fall in the categories of system learning and system mining. The state series of an evolving system is denoted as $SS = \{S1, S2, \ldots, SN\}$. Then, the connections (or relationships) between entities of each state are pre-processed to make a temporal network, and this resulted in a series of evolving networks $EN = \{EN1, EN2, \ldots, ENN\}$. These temporal networks can be merged to make an evolution representor, which is used with learning and mining techniques for system evolution analysis. This made us to analyze evolving inter-connected entities of a system state series. The system learning is performed by applying active learning and deep learning on the evolution representor. The system mining is performed by applying two proposed pattern-mining techniques: network rule mining and subgraph mining. Specifically, the publications describe the following proposed approaches: System State Complexity, Evolving System Complexity, System Evolution Recommender, Stable Network Evolution Rule, and System Changeability Metric. The proposed approaches are used to generate recommendation and evolution information to perform system evolution analysis. For example, a graph theory application of a service change classifier algorithm assigning change labels to a web service's call graph representing calls between operations and procedures, which helped to do Web Service Slicing by extracting a WSDL slice for Inter-operational analysis.

Participants

- José Aliste-Prieto
Universidad Andres Bello – Santiago de Chile, CL
- Nantel Bergeron
York University – Toronto, CA
- Cornelius Brand
Universität des Saarlandes, DE
- Animesh Chaturvedi
Indian Institute of Technology – Indore, IN
- Carolyn Chun
U.S. Naval Academy – Annapolis, US
- Anna De Mier
UPC – Barcelona, ES
- Jo Ellis-Monaghan
Saint Michael's College – Colchester, US
- Graham Farr
Monash University – Clayton, AU
- Alex Fink
Queen Mary University of London, GB
- Delia Garijo
University of Sevilla, ES
- Daniela Genova
University of North Florida – Jacksonville, US
- Emeric Gioan
University of Montpellier & CNRS, FR
- Chris Godsil
University of Waterloo, CA
- Andrew Goodall
Charles University – Prague, CZ
- Krystal Guo
UL – Brussels, BE
- Orli Herscovici
Technion – Haifa, IL
- Hendrik Jan Hoogeboom
Leiden University, NL
- Benjamin Jones
Monash University – Clayton, AU
- Nataša Jonoska
University of South Florida – Tampa, US
- Louis H. Kauffman
Novosibirsk State University, RU
- Martin Kochol
Bratislava, SK
- Thomas Krajewski
Aix-Marseille Université, FR
- Joseph Kung
University of North Texas – Denton, US
- Sergei Lando
NRU Higher School of Economics – Moscow, RU
- Bodo Lass
University Claude Bernard – Lyon, FR
- Johann A. Makowsky
Technion – Haifa, IL
- Iain Moffatt
Royal Holloway University of London, GB
- Kerri Morgan
Deakin University – Melbourne, AU
- Steven Noble
Birkbeck, University of London, GB
- Marc Noy
UPC – Barcelona, ES
- Jaeseong Oh
Seoul National University, KR
- James Oxley
Louisiana State University – Baton Rouge, US
- Vsevolod Rakita
Technion – Haifa, IL
- Elena V. Ravve
ORT Braude College – Karmiel, IL
- Guus Regts
University of Amsterdam, NL
- Adrian Tanasa
University of Bordeaux, FR
- Maya Thompson
Royal Holloway University of London, GB
- Peter Tittmann
Hochschule Mittweida, DE
- Lluis Vena Cros
Charles University – Prague, CZ
- William Whistler
Durham University, GB
- José Zamora Ponce
Universidad Andres Bello – Santiago de Chile, CL