DAGSTUHL
REPORTS

Volume 10, Issue 2, February 2020

Dagstuhl Reports, Vol. 10, Issue 2 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
September, 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.10.2.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.

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 20061

SAT and Interactions

Edited by

Olaf Beyersdorff[1], Uwe Egly[2], Meena Mahajan[3], and Cláudia Nalon[4]

1 **Universität Jena, DE,** olaf.beyersdorff@uni-jena.de
2 **TU Wien, AT,** uwe.egly@tuwien.ac.at
3 **Institute of Mathematical Sciences – Chennai, IN,** meena@imsc.res.in
4 **University of Brasilia, BR,** nalon@unb.br

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 20061 "SAT and Interactions". The seminar brought together theoreticians and practitioners from the areas of proof complexity and proof theory, SAT and QBF solving, MaxSAT, and modal logics, who discussed recent developments in their fields and embarked on an interdisciplinary exchange of ideas and techniques between these neighbouring subfields of SAT.

Seminar February 2–7, 2020 – http://www.dagstuhl.de/20061
2012 ACM Subject Classification Theory of computation → Proof theory, Theory of computation → Modal and temporal logics, Theory of computation → Automated reasoning, Theory of computation → Description logics, Theory of computation → Proof complexity, Theory of computation → Complexity theory and logic
Keywords and phrases SAT, MaxSAT, QBF, proof complexity, deep inference, modal logic, solving
Digital Object Identifier 10.4230/DagRep.10.2.1
Edited in cooperation with Joshua Lewis Blinkhorn

1 Executive Summary

Olaf Beyersdorff
Uwe Egly
Meena Mahajan
Cláudia Nalon

The problem of deciding whether a propositional formula is satisfiable (SAT) is one of the most fundamental problems in computer science, both theoretically and practically. Its theoretical significance derives from the Cook-Levin Theorem, identifying SAT as the first NP-complete problem. Since then SAT has become a reference for an enormous variety of complexity statements, among them the celebrated P vs NP problem: one of seven million-dollar Clay Millennium Problems. Due to its NP hardness, SAT has been classically perceived as an intractable problem, and indeed, unless $\mathsf{P} = \mathsf{NP}$, no polynomial-time algorithm for SAT exists.

There are many generalisations of the SAT problem to further logics, including quantified Boolean formulas (QBFs) and modal and temporal logics. These logics present even harder satisfiability problems as they are associated with complexity classes such as PSPACE, which encompasses NP. However, QBFs, modal and temporal logics can express many practically relevant problems far more succinctly, thus applying to more real-world problems from artificial intelligence, bioinformatics, verification, and planning.

Due to its practical implications, intensive research has been performed on how to solve SAT problems in an automated fashion. The last decade has seen the development of practically efficient algorithms for SAT, QBFs and further logics and their implementation as solvers, which successfully solve huge industrial instances.

Very often, these developments take place within different communities, e.g., there has been almost no interaction between the areas of SAT/QBF solving and solving for modal logics.

The main aim of the proposed Dagstuhl Seminar therefore was to bring together researchers from proof complexity and proof theory, SAT, MaxSAT and QBF solving, and modal logics so that they can communicate state-of-the-art advances and embark on a systematic interaction that will enhance the synergy between the different areas. As such the seminar was the first workshop (in Dagstuhl and elsewhere) to unite researchers working on both theory and practice of propositional SAT, QBF, and modal logics. One of the specific aims was to foster more interaction between these different communities with the goal to transfer the success of theoretical research on SAT to further logics and SAT problems.

To facilitate such interactions, the seminar included a number of survey talks to introduce neighbouring communities to the main notions, results, and challenges of the represented areas. The following survey talks were given during the seminar:

- Massimo Lauria: Proof Complexity: A Survey,
- Lutz Straßburger: Introduction to Deep Inference,
- Vijay Ganesh: Machine Learning and Logic Solvers: The Next Frontier,
- Mikoláš Janota: QBF Solving and Calculi: An Overview,
- João Marques-Silva: Practical MaxSAT Solving: A Survey,
- Cláudia Nalon: Modal Logics: An Overview.

Each of these surveys was accompanied by one or more sessions with contributed talks dedicated to recent specific results of the field.

The seminar also included an open discussion session on ‘Future Directions of Research’, where ideas for a closer interaction between theoretical fields such as proof theory and proof complexity and practical fields such as SAT/QBF and modal solving were discussed.

The organisers believe that the seminar fulfilled their original high goals: most talks were a great success and many participants reported about the inspiring seminar atmosphere, fruitful interactions, and a generally positive experience. The organisers and participants wish to thank the staff and the management of Schloss Dagstuhl for their assistance and excellent support in the arrangement of a very successful and productive event.

2 Table of Contents

Executive Summary

Overview of Talks

Future Directions of Research

3 Overview of Talks

3.1 Hardness Characterisations and Size-Width in QBF Resolution

Joshua Lewis Blinkhorn (Universität Jena, DE), Olaf Beyersdorff (Universität Jena, DE), and Meena Mahajan (Institute of Mathematical Sciences – Chennai, IN)

Main reference Olaf Beyersdorff, Joshua Blinkhorn, Meena Mahajan: "Hardness Characterisations and Size-Width Lower Bounds for QBF Resolution", in Proc. of the 35th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS '20, p. 209–223, Association for Computing Machinery, 2020.
URL https://doi.org/10.1145/3373718.3394793

We provide a tight characterisation of proof size in resolution for quantified Boolean formulas (QBF) by circuit complexity. Such a characterisation was previously obtained for a hierarchy of QBF Frege systems [1], but leaving open the most important case of QBF resolution. Different from the Frege case, our characterisation uses a new version of decision lists as its circuit model, which is stronger than the CNFs the system works with. Our decision list model is well suited to compute countermodels for QBFs. Our characterisation works for both Q-Resolution and QU-Resolution, which we show to be polynomially equivalent for QBFs of bounded quantifier alternation.

Using our characterisation we obtain a size-width relation for QBF resolution in the spirit of the celebrated result for propositional resolution [2]. However, our result is not just a replication of the propositional relation – intriguingly ruled out for QBF in previous research [3] – but shows a different dependence between size, width, and quantifier complexity.

References

1 O. Beyersdorff and J. Pich. Understanding Gentzen and Frege systems for QBF. In: *Symposium on Logic in Computer Science (LiCS)*, pp. 146–155, ACM, 2016.

2 E. Ben-Sasson and A. Wigderson. Short proofs are narrow – resolution made simple. *Journal of the ACM*, 48(2):149–169, 2001.

3 O. Beyersdorff, L. Chew, M. Mahajan, and A. Shukla. Are short proofs narrow? QBF resolution is *not* so simple. *ACM Transactions on Computational Logic*, 19(1):1–26, 2018.

3.2 Tractable QBF via Knowledge Compilation

Florent Capelli (Lille I University, FR)

Joint work of Florent Capelli, Stefan Mengel

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 [1] 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.

References

1 H. Chen. Quantified constraint satisfaction and bounded treewidth. In: *European Conference on Artificial Intelligence (ECAI)*, pp. 161–165, IOS Press, 2004.

3.3 The Equivalences of Refutational QRAT

Leroy Nicholas Chew (Carnegie Mellon University – Pittsburgh, US) and Judith Clymo (University of Leeds, GB)

The solving of Quantified Boolean Formulas (QBF) has been advanced considerably in the last two decades. In response to this, several proof systems have been put forward to universally verify QBF solvers. QRAT by Heule et al. is one such example of this and builds on technology from DRAT, a checking format used in propositional logic. Recent advances have shown conditional optimality results for QBF systems that use extension variables. Since QRAT can simulate Extended Q-Resolution, we know it is strong, but we do not know if QRAT has the strategy extraction property as Extended Q-Resolution does. In this paper, we partially answer this question by showing that a simple restriction on the reduction rule in QRAT is enough to show it has strategy extraction (and consequentially is equivalent to Extended Q-Resolution modulo NP). We also extend equivalence to another system, as we show an augmented version of QRAT known as QRAT+, developed by Lonsing and Egly, is in fact equivalent to the basic QRAT. We achieve this by constructing a line-wise simulation of QRAT+ using only steps valid in QRAT.

3.4 How QBF Expansion Makes Strategy Extraction Hard

Judith Clymo (University of Leeds, GB) and Leroy Nicholas Chew (Carnegie Mellon University – Pittsburgh, US)

Main reference Leroy Chew, Judith Clymo: "How QBF Expansion Makes Strategy Extraction Hard", Electronic Colloquium on Computational Complexity (ECCC), Vol. 26, p. 157, 2019.
URL https://eccc.weizmann.ac.il/report/2019/157

In this talk we show that the QBF proof checking format QRAT (Quantified Resolution Asymmetric Tautologies) by Heule, Biere and Seidl cannot have polynomial-time strategy extraction unless P=PSPACE. In our proof, the crucial property that makes strategy extraction PSPACE-hard for this proof format is universal expansion, even expansion on a single variable.

While expansion reasoning used in other QBF calculi can admit polynomial time strategy extraction, we find this is conditional on a property studied in proof complexity theory. We show that strategy extraction on expansion based systems can only happen when the underlying propositional calculus has the property of feasible interpolation.

3.5 From QBFs to MALL and Back via Focussing

Anupam Das (University of Birmingham, GB)

In this work we investigate how to extract alternating time bounds from "focussed" proof systems. Our main result is the obtention of fragments of MALLw (MALL with weakening) complete for each level of the polynomial hierarchy. In one direction we encode QBF satisfiability and in the other we encode focussed proof search, and we show that the composition of the two encodings preserves quantifier alternation, yielding the required result. By carefully composing with well-known embeddings of MALLw into MALL, we obtain a similar delineation of MALL formulas, again carving out fragments complete for each level of the polynomial hierarchy. This refines the well-known results that both MALLw and MALL are PSPACE-complete.

A key insight is that we have to refine the usual presentation of focussing to account for deterministic computations in proof search, which correspond to invertible rules that do not branch. This is so that we may more faithfully associate phases of focussed proof search to their alternating time complexity. This presentation seems to uncover further dualities at the level of proof search than usual presentations, so could be of further proof theoretic interest in its own right.

3.6 Consistent Query Answering via SAT Solving

Akhil Dixit (University of California – Santa Cruz, US)

Consistent Query Answering is a rigorous and principled approach to the semantics of queries posed against inconsistent databases, i.e., the databases that violate one or more integrity constraints set over its schema. Computing the consistent answers to a fixed conjunctive query on a given inconsistent database can be a coNP-hard problem, even though every fixed conjunctive query is efficiently computable on a given consistent database. In this talk, we will first introduce some database problems, their connections to SAT, and some recent theoretical results in the literature. In the later half, we will present CAvSAT (Consistent Answers via SAT), our SAT-based system for consistent query answering.

3.7 Clausal Resolution for Temporal Logics

Clare Dixon (University of Liverpool, GB)

Joint work of Michael Fisher, Clare Dixon, Martin Peim
Main reference Michael Fisher, Clare Dixon, Martin Peim: "Clausal Temporal Resolution", ACM Trans. Comput. Logic, Vol. 2(1), p. 12–56, Association for Computing Machinery, 2001.
URL https://doi.org/10.1145/371282.371311

A clausal temporal resolution calculus has been developed [1] and implemented [2] for linear-time temporal logics (LTL). This involves translation to a clausal normal form, resolution rules that apply to formulae holding at the same time moment and a loop resolution rule that applies across time moments. This approach has been extended to other temporal (and modal [6, 5]) logics such as the branching time temporal logic CTL [7] and first-order temporal logic [3, 4]. We discuss the main elements of this approach applied to LTL and its extensions to other non-classical logics.

References

1 M. Fisher, C. Dixon, and M. Peim. Clausal temporal resolution. *ACM Transactions on Computational Logic*, 2(1):12–56, 2001.

2 U. Hustadt and Boris Konev. TRP++ 2.0: A temporal resolution prover. In: *Conference on Automated Deduction* (*CADE*), pp. 274–278. Springer, 2003.

3 B. Konev, A. Degtyarev, C. Dixon, M. Fisher, and U. Hustadt. Mechanising first-order temporal resolution. *Information and Computation*, 199(1-2):55–86, 2005.

4 M. Ludwig and U. Hustadt. Implementing a fair monodic temporal prover. *AI Communications*, 23(2-3):68–96, 2010.

5 C. Nalon, U. Hustadt, and C. Dixon. KSP: A resolution-based theorem prover for K_n: architecture, refinements, strategies and experiments. *Journal of Automated Reasoning*, 64(3):461–484, 2020.

6 Cláudia Nalon, Clare Dixon, and Ullrich Hustadt. Modal resolution: proofs, layers, and refinements. *ACM Transactions on Computational Logic*, 20(4):23:1–23:38, 2019.

7 L. Zhang, U. Hustadt, and C. Dixon. A resolution calculus for the branching-time temporal logic CTL. *ACM Transactions on Computational Logic*, 15(1):1529–3785, 2014.

3.8 Machine Learning and Logic Solving: the Next Frontier

Vijay Ganesh (University of Waterloo, CA)

Over the last two decades, software engineering has witnessed a silent revolution in the form of Boolean SAT solvers. These solvers are now integral to many analysis, synthesis, verification, and testing approaches. This is largely due to a dramatic improvement in the scalability of these solvers vis-a-vis large real-world formulas. What is surprising is that the Boolean satisfiability problem is NP-complete, believed to be intractable in general, and yet these solvers easily solve instances containing millions of variables and clauses in them. How can that be?

In my talk, I will address this question of why SAT solvers are so efficient through the lens of machine learning as well as ideas from (parameterized) proof complexity. I will argue that SAT solvers are best viewed as proof systems, composed of prediction engines that optimize some metric correlated with solver running time. These prediction engines can be built using ML techniques, whose aim is to structure solver proofs in an optimal way. Thus, two major paradigms of AI, namely machine learning and logical deduction, are brought together in a principled way to design efficient SAT solvers. A result of my research is the MapleSAT solver, that has been the winner of several recent international SAT competitions, and is now widely used in industry and academia.

3.9 Semi-Algebraic Proofs, IPS Lower Bounds and the τ-Conjecture: Can a Natural Number be Negative?

Edward A. Hirsch (Steklov Institute – St. Petersburg, RU)

Joint work of Yaroslav Alekseev, Dima Grigoriev, Edward A. Hirsch, Iddo Tzameret
Main reference Yaroslav Alekseev, Dima Grigoriev, Edward A. Hirsch, Iddo Tzameret: "Semi-Algebraic Proofs, IPS Lower Bounds, and the τ-Conjecture: Can a Natural Number Be Negative?", in Proc. of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, p. 54–67, Association for Computing Machinery, 2020.
URL https://doi.org/10.1145/3357713.3384245

We show that the equivalence between algebraic and semialgebraic proofs represented by circuits is tightly connected to refuting the bit value principle (BVP): $\sum x_i 2^i = -1$. We relate the complexity of BVP to previously known conjectures in algebraic complexity.

3.10 Benchmarking Modal Logic Theorem Provers

Ullrich Hustadt (University of Liverpool, GB)

Modal logics are extensions of propositional and first-order logic with operators that are not truth-functional. The contemporary era of modal logics started in the late 50s and development of theorem provers started in earnest in the late 80s. Since then there has been an interest in the evaluation of the practical performance of such provers. I will discuss some of the approaches that have been used for such evaluations over the past thirty years.

3.11 QBF Solving and Calculi: An Overview

Mikoláš Janota (IST – Lisbon, PT)

Quantified Boolean Formulas (QBFs) enrich the SAT problem with quantifiers taking the problem from NP to PSPACE. Recent years have seen a number of novel approaches to QBF solving. At the same time, QBF calculi were developed to match the solvers. However, there are calculi with no solving counterparts.

In this talk I will overview the two prominent paradigms in QBF solving: conflict-driven and expansion-based. I will also discuss the connection between solving and the existing proof systems as well as challenges for future research.

3.12 Simplified and Improved Separation Between Regular and General Resolution by Lifting

Jan Johannsen (LMU München, DE)

Joint work of Marc Vinyals, Jan Elffers, Jan Johannsen, Jakob Nordström
Main reference Marc Vinyals, Jan Elffers, Jan Johannsen, Jakob Nordström: "Simplified and improved separations between regular and general resolution by lifting". 2020 (submitted for publication).

We give a significantly simplified proof of the exponential separation between regular and general resolution of Alekhnovich et al. (2007) as a general theorem lifting proof depth to regular proof length in resolution. This simpler proof then allows us to strengthen the separation further, and to construct families of theoretically very easy benchmarks that are surprisingly hard for SAT solvers in practice.

3.13 Proof Complexity: A Survey

Massimo Lauria (Sapienza University of Rome, IT)

Running a SAT solver on an UNSAT formula produces (implicitly or explicitly) a proof that the formula is unsatisfiable, usually expressible in an established formal language called a proof system. This observation leads to study SAT solving methods by looking at the generated proofs. That is, if an UNSAT formula has no short proof in a given proof system, the corresponding SAT solvers cannot solve the formula efficiently.

We introduce the area and the methods of proof complexity, that studies the length and the structure of proofs of UNSAT. In particular we define and discuss Resolution, Polynomial Calculus, Cutting Planes and DRAT/Extended Resolution, which are the most relevant proof systems for current SAT solver technology.

Since a SAT solver's goal is, among other things, to look for proofs as short as possible, we briefly discuss the complexity of efficiently finding such short proofs.

3.14 Practical MaxSAT Solving: A Survey

João Marques-Silva (University of Toulouse, FR)

This talk presents an overview of practical algorithms for maximum satisfiability (MaxSAT) solving, highlighting the algorithms that are most effective in practice.

3.15 Revisiting Graph Width Measures for CNF-Encodings

Stefan Mengel (CNRS, CRIL – Lens FR)

We consider bounded width CNF-formulas where the width is measured by popular graph width measures on graphs associated to CNF-formulas. Such restricted graph classes, in particular those of bounded treewidth, have been extensively studied for their uses in the design of algorithms for various computational problems on CNF-formulas. Here we consider the expressivity of these formulas in the model of clausal encodings with auxiliary variables. We first show that bounding the width for many of the measures from the literature leads to a dramatic loss of expressivity, restricting the formulas to those of low communication complexity. We then show that the width of optimal encodings with respect to different measures is strongly linked: there are two classes of width measures, one containing primal treewidth and the other incidence cliquewidth, such that in each class the width of optimal encodings only differs by constant factors. Moreover, between the two classes the width differs at most by a factor logarithmic in the number of variables. Both these results are in stark contrast to the setting without auxiliary variables where all width measures we consider here differ by more than constant factors and in many cases even by linear factors.

3.16 Modal Logics: An Overview (Parts I and II)

Cláudia Nalon (University of Brasilia, BR)

Joint work of Clare Dixon, Ullrich Hustadt

The first talk was a gentle introduction to modal logics, focusing on the basic multimodal logic K_n. We have discussed the different reasoning tasks for this class of logics, local and global, and their complexity. We have also introduced two different calculi for dealing with reasoning in modal settings: tableaux and resolution.

The second talk was dedicated to the layered resolution calculus for local reasoning for K_n. The calculus is inspired by model-theoretical results concerning satisfiability in K_n: models can be restricted to finite tree-like structures and the satisfiability of a (sub)formula depends only of the subtree in which such a subformula occurs. Clauses are labelled by the height (the modal depth) they occur in such a tree. Inference rules can only be applied to clauses whose labels unify. This restricts the candidates for resolution whilst retaining completeness. Experimental results show that the theorem-prover which implements the calculus, KSP, works well on problems with high modal depth and uniform distribution of propositional symbols over the different depths.

References

1 Cláudia Nalon, Clare Dixon, and Ullrich Hustadt. Modal resolution: Proofs, layers, and refinements. *ACM Transactions on Computational Logic*, 20(4):23:1–23:38, 2019.

2 C. Nalon, U. Hustadt, and C. Dixon. KSP: A resolution-based theorem prover for K_n: architecture, refinements, strategies and experiments. *Journal of Automated Reasoning*, 64(3):461–484, 2020.

3.17 Proof and Refutation: An Adventure in Formalisation

Dirk Pattinson (Australian National University – Canberra, AU)

In this talk, we take the point of view that a simple 'yes/no' answer is not a sufficient output of an automated reasoning procedure or implementation. On top of the answer, we demand verifiable evidence, either for provability or refutability of a formula. Clearly, a (formal) proof satisfies this requirement in the case of a provable formula. Countermodels can be used to give evidence of non-provability, but suffer drawbacks: First, there may not be an agreed upon notion of semantics, and second, the mathematical details of countermodels vary widely depending on the underlying logic, while proofs have a very uniform representation.

For this reason, we complement the syntactic notion of proof with a syntactic (coinductively defined) notion of refutation. Our main theorem then states that 'every sequent either has a proof or a refutation' (terms and conditions apply). We speak both on the notion of refutation in general, as well as highlight the challenges encountered in fully verifying the above theorem.

3.18 Dependency Learning for QBF

Tomáš Peitl (Universität Jena, DE), Friedrich Slivovsky (TU Wien, AT), and Stefan Szeider

Main reference Tomás Peitl, Friedrich Slivovsky, Stefan Szeider: "Dependency Learning for QBF", J. Artif. Intell. Res., Vol. 65, pp. 180–208, 2019.
URL http://dx.doi.org/10.1613/jair.1.11529

Quantified Boolean Formulas (QBFs) can be used to succinctly encode problems from domains such as formal verification, planning, and synthesis. One of the main approaches to QBF solving is Quantified Conflict Driven Clause Learning (QCDCL). By default, QCDCL assigns variables in the order of their appearance in the quantifier prefix so as to account for dependencies among variables. Dependency schemes can be used to relax this restriction and exploit independence among variables in certain cases, but only at the cost of nontrivial interferences with the proof system underlying QCDCL.

We introduce dependency learning, a new technique for exploiting variable independence within QCDCL that allows solvers to learn variable dependencies on the fly. The resulting version of QCDCL enjoys improved propagation and increased flexibility in choosing variables for branching while retaining ordinary (long-distance) Q-resolution as its underlying proof system. We show that dependency learning can achieve exponential speedups over ordinary QCDCL. Experiments on standard benchmark sets demonstrate the effectiveness of this technique.

3.19 Forgetting-Based Ontology Extraction

Renate Schmidt (University of Manchester, GB)

Joint work of Jieying Chen, Ghadah Alghamdi, Renate A. Schmidt, Dirk Walther, Yongsheng Gao
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 https://doi.org/10.1145/3360901.3364424

Ontology extraction is an essential operation for the reuse, creation, evaluation, curation, decomposition, integration and general use of ontologies. A method with higher precision is forgetting, also known as uniform interpolation. Forgetting creates a compact representation of a part of the information contained in an ontology that preserves the underlying logical definitions of the specified terms (the interpolation signature) by hiding the remaining terms. This allows users to focus exactly on the information they are interested in.

After an introduction of the idea of forgetting in contrast to modularisation, another ontology extraction method, the presentation gave an overview of recent and current research on developing practical forgetting tools for description logic based ontologies. We also discussed the application of these tools to SNOMED CT, a comprehensive ontology of standardised medical content used in health care systems across several countries.

3.20 Spinal Atomic Lambda-Calculus

David R. Sherratt (Universität Jena, DE)

Joint work of Heijltjes, Willem; Gundersen, Tom; Parigot, Michel
Main reference David Sherratt: "A lambda-calculus that achieves full laziness with spine duplication", 2019.
URL http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.787527

We present the spinal atomic lambda-calculus, a typed lambda-calculus with explicit sharing and atomic duplication that achieves spinal full laziness: duplicating only the direct paths between a binder and bound variables is enough for beta reduction to proceed. We show this calculus is the result of a Curry–Howard style interpretation of a deep-inference proof system, and prove that it has natural properties with respect to the lambda-calculus: confluence and preservation of strong normalisation.

References

1 David Rhys Sherratt. *A Lambda-Calculus that achieves full laziness with spine duplication*. PhD thesis, University of Bath, UK, 2019.

3.21 Computing Unique Strategy Functions by Interpolation

Friedrich Slivovsky (TU Wien, AT)

We present a new semantic gate extraction technique for propositional formulas based on interpolation. While known gate detection methods are incomplete and rely on pattern matching or simple semantic conditions, this approach can detect any definition entailed by an input formula. As an application, we consider the problem of computing unique strategy functions from Quantified Boolean Formulas (QBFs) and Dependency Quantified Boolean Formulas (DQBFs). Experiments with a prototype implementation demonstrate that functions can be efficiently extracted from formulas in standard benchmark sets, and that many of these definitions remain undetected by syntactic gate detection. We turn this into a preprocessing technique by substituting unique strategy functions for input variables and test solver performance on the resulting instances. Compared to syntactic gate detection, we see a significant increase in the number of solved 2QBF instances, as well as modest increases for general QBF and DQBF.

3.22 Introduction to Deep Inference

Lutz Straßburger (INRIA Saclay – Île-de-France, FR)

In the first half of the talk I gave a rough overview over the main research results of deep inference of the last 20 years. I discussed *atomicity* and *locality* of inference rules, I presented proof systems for classical logic and linear logic, and I showed how cut elimination can be proved using the techniques of *decomposition* and *splitting*. A good starting point for reading about this subject are the lecture notes for a course on deep inference given at ESSLLI 2019 (to be found at https://hal.inria.fr/hal-02390267). More information on deep inference can be found on the *deep inference webpage* (http://alessio.guglielmi.name/res/cos/index.html) maintained by Alessio Guglielmi.

In the second half of the talk I discussed proof systems for intuitionistic modal logics using nested sequents. I also presented the prover MOIN, for which a system description is available at https://hal.inria.fr/hal-02457240, where also more references on nested sequent proof systems can be found.

3.23 Hard Examples for Common Variable Decision Heuristics

Marc Vinyals (Technion – Haifa, IL)

Main reference Marc Vinyals: "Hard Examples for Common Variable Decision Heuristics", in Proc. of the The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pp. 1652–1659, AAAI Press, 2020.
URL https://aaai.org/ojs/index.php/AAAI/article/view/5527

The CDCL algorithm, which is nowadays the top-performing algorithm to solve SAT in practice, is polynomially equivalent to resolution when we view it as a proof system, that is we replace some of its heuristics by nondeterministic choices.

In this talk we show that this is no longer true if we leave the heuristics in place; more precisely we build a family of formulas that have resolution proofs of polynomial size but require exponential time to decide in CDCL with a class of variable decision heuristics that includes the most common heuristics such as VSIDS.

3.24 Reversible Pebble Games and the Relation Between Tree-Like and General Resolution Space

Florian Wörz (Universität Ulm, DE) and Jacobo Torán (Universität Ulm, DE)

Joint work of Jacobo Torán, Florian Wörz
Main reference Jacobo Torán, Florian Wörz: "Reversible Pebble Games and the Relation Between Tree-Like and General Resolution Space", in Proc. of the 37th International Symposium on Theoretical Aspects of Computer Science, STACS 2020, March 10-13, 2020, Montpellier, France, LIPIcs, Vol. 154, pp. 60:1–60:18, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020.
URL http://dx.doi.org/10.4230/LIPIcs.STACS.2020.60

We show a new connection between the space measure in tree-like resolution and the reversible pebble game in graphs. Using this connection we provide several formula classes for which there is a logarithmic factor separation between the space complexity measure in tree-like and general resolution. We show that these separations are almost optimal by proving upper bounds for tree-like resolution space in terms of general resolution clause and variable space. In particular we show that for any formula F, its tree-like resolution is upper bounded by $\mathrm{space}(\pi) \log \mathrm{time}(\pi)$, where π is any general resolution refutation of F. This holds considering as $\mathrm{space}(\pi)$ the clause space of the refutation as well as considering its variable space. For the concrete case of Tseitin formulas we are able to improve this bound to the optimal bound $\mathrm{space}(\pi) \log n$, where n is the number of vertices of the corresponding graph.

4 Future Directions of Research

The seminar featured a session in which participants were invited to give a short informal presentation on the theme *directions of future research*. Five partipants contributed presentations, summarised below.

4.1 Anupam Das

The community might like to explore what proof theory can give to solving, because techniques developed in the context of proof theory could have a much wider applicability. For instance, some solvers dealing with the logic K5 translate the formulas to QBF, while preserving some semantic meaning. Performance is heavily dependent on the encoding used. Instead, could one use proof-theoretic techniques like focussing, deep inference, or graph-based concepts? As long as these techniques preserve some measures of the formulas, it may be possible to transfer algorithmic bounds. A somewhat related question would ask what the appropriate measures may be.

4.2 Vijay Ganesh

Proof complexity theorists usually focus on proving lower bounds for various proof systems. However, we have a very poor understanding of why solvers perform well in practice. One approach to deepen this understanding is to establish some parametric upper bounds. The choice of parameters for such a study should be informed by practice. Analysing a solver directly may be hard, but it may be easier to establish a polynomial equivalence to algorithms that establish such parametric upper bounds.

An example in another domain, where theoretical tools give a good explanation for why an algorithm that has bad worst-case behaviour nonetheless behaves extremely well in practice, is the smoothed analysis framework applied to the Simplex algorithm.

4.3 Ullrich Hustadt

There exists a plethora of logics – modal logics, description logics, and so on – each with its own calculi. A typical goal is to extend the reach of Resolution to these logics. As the calculi often arise from philosophy, it is not clear a priori that they all merit such a study. Therefore, it is natural to consider which logics amongst this huge zoo are really worth exploring. The community may then focus its attention there, since it is not large enough to explore them all.

4.4 Oliver Kullmann

Despite great effort, the use of SAT solvers – particularly in the solution of hard problems via schemes like cube-and-conquer – remains poorly understood. An informal "psychoanalysis" could partition the workflow into three phases:

1. the entirely predictable part, based on mathematical truths,
2. the intermediate part, lookahead using statistical tools,
3. the almost utterly unpredictable part – CDCL solvers lie here, and there seems to be some unstable chaos.

Restrictions by partial assignments take us outwards, from families that can be handled within phase 1 alone, to versions that are more chaotic or unstable. How can we explain this phenomenon and put it to use?

4.5 Lutz Straßburger

There are many very different proof systems, and proofs in a specific system are heavily tied to the specifics of the proof system. And yet, (almost) all proofs rely on a common mathematical underpinning. Can we seek proofs independent of the proof system? That is, can we speak of proofs independent of representations in specific proof systems? We want to somehow pinpoint the essential mathematical content of a proof, which will of course depend on the logic, but can it be independent of the proof system? For classical propositional logic, for example, combinatorial proofs is a candidate approach.

Participants

- Olaf Beyersdorff
Universität Jena, DE
- Joshua Lewis Blinkhorn
Universität Jena, DE
- Benjamin Böhm
Universität Jena, DE
- Ilario Bonacina
UPC Barcelona Tech, ES
- Florent Capelli
Lille I University, FR
- Leroy Nicholas Chew
Carnegie Mellon University – Pittsburgh, US
- Judith Clymo
University of Leeds, GB
- Nadia Creignou
Aix-Marseille University, FR
- Anupam Das
University of Birmingham, GB
- Susanna de Rezende
The Czech Academy of Sciences – Prague, CZ
- Akhil Dixit
University of California – Santa Cruz, US
- Clare Dixon
University of Liverpool, GB
- Uwe Egly
TU Wien, AT
- Vijay Ganesh
University of Waterloo, CA
- Azza Gaysin
Charles University – Prague, CZ
- Edward A. Hirsch
Steklov Institute – St. Petersburg, RU
- Ullrich Hustadt
University of Liverpool, GB
- Mikoláš Janota
IST – Lisbon, PT
- Jan Johannsen
LMU München, DE
- Hans Kleine Büning
Universität Paderborn, DE
- Oliver Kullmann
Swansea University, GB
- Massimo Lauria
Sapienza University of Rome, IT
- Meena Mahajan
Institute of Mathematical Sciences – Chennai, IN
- Joao Marques-Silva
University of Toulouse, FR
- Barnaby Martin
Durham University, GB
- Stefan Mengel
CNRS, CRIL – Lens FR
- Claudia Nalon
University of Brasilia, BR
- Jakob Nordström
University of Copenhagen, DK & Lund University, SE
- Dirk Pattinson
Australian National University – Canberra, AU
- Tomáš Peitl
Universität Jena, DE
- Renate Schmidt
University of Manchester, GB
- Uwe Schöning
Universität Ulm, DE
- David R. Sherratt
Universität Jena, DE
- Anil Shukla
Indian Institute of Technology Ropar – Rupnagar, IN
- Friedrich Slivovsky
TU Wien, AT
- Gaurav Sood
Institute of Mathematical Sciences – Chennai, IN
- Lutz Straßburger
INRIA Saclay – Île-de-France, FR
- Jacobo Torán
Universität Ulm, DE
- Marc Vinyals
Technion – Haifa, IL
- Florian Wörz
Universität Ulm, DE

Report from Dagstuhl Seminar 20071

Foundations of Composite Event Recognition

Edited by

Alexander Artikis[1], Thomas Eiter[2], Alessandro Margara[3], and Stijn Vansummeren[4]

1 **University of Piraeus, GR & NCSR Demokritos, GR,** a.artikis@iit.demokritos.gr
2 **TU Wien, AT,** eiter@kr.tuwien.ac.at
3 **Polytechnic University of Milan, IT,** alessandro.margara@polimi.it
4 **Université Libre de Bruxelles, BE,** stijn.vansummeren@ulb.ac.be

Abstract

Composite Event Recognition (CER) refers to the activity of detecting patterns in streams of continuously arriving "event" data over, possibly geographically, distributed sources. CER is key in Big Data applications that require the processing of such event streams to obtain timely insights and to implement reactive and proactive measures. Examples include the recognition of emerging stories and trends on the Social Web, traffic and transport incidents in smart cities, and epidemic spread.

Numerous CER languages have been proposed in the literature. While these systems have a common goal, they differ in their data models, pattern languages and processing mechanisms, resulting in heterogeneous implementations with fundamentally different capabilities. Moreover, we lack a common understanding of the trade-offs between expressiveness and complexity, and a theory for comparing the fundamental capabilities of CER systems. As such, CER frameworks are difficult to understand, extend and generalise. It is unclear which of the proposed approaches better meets the requirements of a given application. Furthermore, the lack of foundations makes it hard to leverage established results – from automata theory, temporal logics, etc – thus hindering scientific and technological progress in CER.

The objective of the seminar was to bring together researchers and practitioners working in Databases, Distributed Systems, Automata Theory, Logic and Stream Reasoning; disseminate the recent foundational results across these fields; establish new research collaborations among these fields; thereby start making progress towards formulating such foundations.

Seminar February 9–14, 2020 – http://www.dagstuhl.de/20071

2012 ACM Subject Classification Computing methodologies → Knowledge representation and reasoning, Theory of computation, Information systems → Data management systems, Information systems → Data streaming

Keywords and phrases complex event processing, event algebra, pattern matching, stream reasoning, temporal reasoning

Digital Object Identifier 10.4230/DagRep.10.2.19

Edited in cooperation with Elias Alevizos

Foundations of Composite Event Recognition, *Dagstuhl Reports*, Vol. 10, Issue 2, pp. 19–49
Editors: Alexander Artikis, Thomas Eiter, Alessandro Margara, and Stijn Vansummeren
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Alessandro Margara (Polytechnic University of Milan, IT)
Alexander Artikis (University of Piraeus, GR & NCSR Demokritos, GR)
Thomas Eiter (TU Wien, AT)
Stijn Vansummeren (Université Libre de Bruxelles, BE)

This report contains the program and outcomes of Dagstuhl Seminar 20071 on "Foundations of Composite Event Recognition" held at Schloss Dagstuhl, Leibniz Center for Informatics, during February 9-14, 2020.

Composite Event Recognition (CER for short) refers to the activity of detecting patterns in streams of continuously arriving "event" data over, possibly geographically, distributed sources. CER is a key ingredient of many contemporary Big Data applications that require the processing of such event streams in order to obtain timely insights and implement reactive and proactive measures. Examples of such applications include the recognition of attacks in computer network nodes, human activities on video content, emerging stories and trends on the Social Web, traffic and transport incidents in smart cities, error conditions in smart energy grids, violations of maritime regulations, cardiac arrhythmia, and epidemic spread. In each application, CER allows to make sense of streaming data, react accordingly, and prepare for counter-measures.

CER systems become increasingly important as we move from an information economy to an "intelligent economy", where it is not only the accessibility to information that matters but also the ability to analyse, reason, and act upon information, creating competitive advantage in commercial transactions, enabling sustainable management of communities, and promoting appropriate distribution of social, healthcare, and educational services. Current businesses tend to be unable to make sense of the amounts of data that are generated by the increasing number of distributed data sources that are becoming available daily, and rely more and more on CER. As an example, traffic management in smart cities requires the analysis of data from an increasing number of sensors, both mobile (mounted on public transport vehicles and private cars) and stationary (installed on intersections). Using such data streams, CER may be used to detect or even forecast traffic congestions, thus allowing for proactively changing traffic light policies and speed limits, with the aim of reducing carbon emissions, optimising public transportation, and improving the quality of life and productivity of commuters. As another example, in smart energy grids, streaming information from power grid elements sensors, end-user devices, and diverse other sources such as weather forecasts and event schedules can be combined through CER to improve the grid efficiency and meet the rapidly increasing electricity demand.

Numerous CER systems and languages have been proposed in the literature. While these systems have a common goal, they differ in their architectures, data models, pattern languages, and processing mechanisms, resulting in many heterogeneous implementations with sometimes fundamentally different capabilities. Their comparative assessment is further hindered by the fact that they have been developed in different communities, each bringing in their own terminology and view of the problem.

Moreover, the established CER literature focuses on the practical system aspects of CER. As a result, little work has been done on its formal foundations. Consequently, and in contrast to the situation for more traditional fields in Computer Science, we currently lack a common understanding of the trade-offs between expressiveness and complexity in the design of CER systems, as well as an established theory for comparing their fundamental capabilities.

As such, currently, CER frameworks are difficult to understand, extend and generalise. It is unclear which of the proposed approaches better meets the requirements of a given application domain, in terms of capturing the intended meaning of the composite events of interest, as well as detecting them efficiently. Furthermore, the lack of foundations makes it hard to leverage established results – from automata theory, temporal logics, etc – thus hindering scientific and technological progress in CER.

At the same time, recent years have witnessed increased activities in diverse fields of Computer Science on topics that are related to CER: Inductive and deductive reasoning over streaming data, a field known as Stream Reasoning in Artificial Intelligence. Theoretical complexity results related to processing database queries under updates, associated with advances in Incremental View Maintenance in Database Research. Expressiveness and complexity of logics in the dynamic setting, in Logic research.

The seminar brought together 39 researchers and practitioners working in domains that are strictly related to CER. The first days of the seminar mainly focused on tutorials and talks that gave an overview of the approaches, techniques, methodologies, and vocabularies used in different communities to refer to CER problems. In particular, the following tutorials were presented:

- Applications and requirements for CER
- CER in data management
- CER in distributed event-based systems
- Stream reasoning
- CER in logic and AI
- CER in business process management

The seminar continued by alternating sessions with focused research talks and group discussions on the following topics, that the participants identified as the most relevant for future investigations and research efforts:

- CER language formalisms
- Towards a common framework for CER expressiveness and complexity
- Evaluation strategies: parallel and distributed processing
- Uncertainty in CER
- Pattern induction and composite event forecasting
- Benchmarking

The final sessions of the seminar focused on reporting the results of the group discussions and in planning follow-up activities, including co-organized workshops and events, joint publications, and projects.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 Complex Event Forecasting

Elias Alevizos (NCSR Demokritos – Athens, GR)

Complex Event Processing (CEP) systems have appeared in abundance during the last two decades. Their purpose is to detect in real-time interesting patterns upon a stream of events and to inform an analyst for the occurrence of such patterns in a timely manner.However, there is a lack of methods for forecasting when a pattern might occur before such an occurrence is actually detected by a CEP engine. We present Wayeb, a framework that attempts to address the issue of Complex Event Forecasting. Wayeb employs symbolic automata as a computational model for pattern detection and variable-order Markov models for deriving a probabilistic description of a symbolic automaton.

References

1 Elias Alevizos, Alexander Artikis, Georgios Paliouras. *Event Forecasting with Pattern Markov Chains*. DEBS, 2017.
2 Elias Alevizos, Alexander Artikis, Georgios Paliouras. *Wayeb: a Tool for Complex Event Forecasting*. LPAR, 2018.

3.2 Stream Logic

François Bry (LMU München, DE)

Joint work of François Bry, Thomas Prokosch

Stream Logic is an attempted formalisation of data streams in predicate logic aimed at enhancing logic programming with streams of (possibly complex) events. The talk motivates Stream Logic with applications, sketches a syntax and a semantics based on a model theory, and relatesStream Logic to meta-programming and non-well-founded sets.

3.3 Complex Event Recognition in Logic and AI: A Tutorial

Diego Calvanese (Free University of Bozen-Bolzano, IT)

In this tutorial we discuss different frameworks, formalisms, and languages, that have been developed within the communities of knowledge representation and reasoning, formal verification, and also database theory and that are in one way or another relevant for the area of Complex Event Recognition (CER). Such formalisms typically rely on combining variants of temporal logics with logics used in knowledge representation and reasoning, and such combination poses challenges with respect to both semantics and computability. The challenges have been addressed by adopting a variety of techniques and by making various assumptions, but the area is still very fragmented and there is no unifying or consolidated framework. The aim of the presentation is to identify opportunities for collaboration and for cross-fertilization with the CER community.

3.4 Whole-system Provenance and Composite Event Recognition

David Eyers (University of Otago, NZ)

Joint work of Thomas F. J.-M. Pasquier, Jatinder Singh, David M. Eyers, Jean Bacon
Main reference Thomas F. J.-M. Pasquier, Jatinder Singh, David M. Eyers, Jean Bacon: "Camflow: Managed Data-Sharing for Cloud Services", IEEE Trans. Cloud Computing, Vol. 5(3), pp. 472–484, 2017.
URL https://doi.org/10.1109/TCC.2015.2489211

Whole-system provenance is becoming increasingly practical as a means to track and audit the way in which data travels throughout computer systems. For example, Pasquier's CamFlow provenance system is implemented as a Linux Security Module so as to facilitate collecting fine-grained provenance data across both kernel and user-space. However, many forms of provenance query risk rapidly generating unmanageably large volumes of logging information, much of which is typically of little value. The CamQuery extension to CamFlow provides means to facilitate run-time, in-kernel data filtering, involving the detection of patterns of interest within the graph data that is generated during provenance tracking. CamQuery already achieves some forms of composite event recognition (CER), but there is significant potential to integrate CER approaches more directly into CamQuery, to ease querying whole-system provenance.

3.5 Distributed Data Streaming and the Power of Geometry

Minos Garofalakis (Technical University of Crete – Chania, GR)

Effective Big Data analytics pose several difficult challenges for modern data management architectures. One key such challenge arises from the naturally streaming nature of big data, which mandates efficient algorithms for querying and analyzing massive, continuous data streams (that is, data that is seen only once and in a fixed order) with limited memory and CPU-time resources. In addition to memory- and time-efficiency concerns, the inherently distributed nature of such applications also raises important communication-efficiency issues, making it critical to carefully optimize the use of the underlying network infrastructure. In this talk, we introduce the distributed data streaming model, and discuss techniques for tracking complex queries over distributed streams that rely on novel insights from convex geometry. We also outline possible research directions in this space.

3.6 Towards streaming evaluation of queries with correlation in complex event processing

Alejandro J. Grez (PUC – Santiago de Chile, CL)

Joint work of Alejandro J. Grez, Riveros, Cristian
Main reference Alejandro Grez, Cristian Riveros, Martín Ugarte: "A Formal Framework for Complex Event Processing", in Proc. of the 22nd International Conference on Database Theory, ICDT 2019, March 26-28, 2019, Lisbon, Portugal, LIPIcs, Vol. 127, pp. 5:1–5:18, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019.
URL https://doi.org/10.4230/LIPIcs.ICDT.2019.5

Complex event processing (CEP) has gained a lot of attention for evaluating complex patterns over high-throughput data streams. Recently, new algorithms for the evaluation of CEP patterns have emerged with strong guarantees of efficiency, i.e. constant update-time per tuple and constant-delay enumeration. Unfortunately, these techniques are restricted for patterns with local filters, limiting the possibility of using joins for correlating the data of events that are far apart.

In this work, we embark on the search for efficient evaluation algorithms of CEP patterns with joins. We start by formalizing the so-called partition-by operator, a standard operator in data stream management systems to correlate contiguous events on streams. Although this operator is a restricted version of a join query, we show that partition-by (without iteration) is equally expressive as hierarchical queries, the biggest class of full conjunctive queries that can be evaluated with constant update-time and constant-delay enumeration over streams. To evaluate queries with partition-by we introduce an automata model, called chain complex event automata (chain-CEA), an extension of complex event automata that can compare data values by using equalities and disequalities. We show that this model admits determinization and is expressive enough to capture queries with partition-by. More importantly, we provide an algorithm with constant update time and constant delay enumeration for evaluating any query definable by chain-CEA, showing that all CEP queries with partition-by can be evaluated with these strong guarantees of efficiency.

3.7 Event Stream Processing with BeepBeep

Sylvain Hallé (University of Quebec at Chicoutimi, CA)

Main reference Sylvain Hallé: "Event Stream Processing with BeepBeep 3: Log Crunching and Analysis Made Easy". Presses de l'Université du Québec, ISBN 978-2-7605-5101-5, 332 pages, 2018.

BeepBeep is simple general purpose event stream processing library. This talk will give a short introduction to the system, and highlight some of its distinguishing features. BeepBeep is based on the concept of *processors*, which are simple, stateful units of computation that can be composed to perform complex processing chains. These chains are created using Java code, but can usually be represented graphically using standardized pictograms, as in Figure 1.

In particular, in this presentation we have shown how it is possible to write Domain-Specific Languages in a few lines of code, and how BeepBeep integrates some elements of explainability and traceability for its computed results.

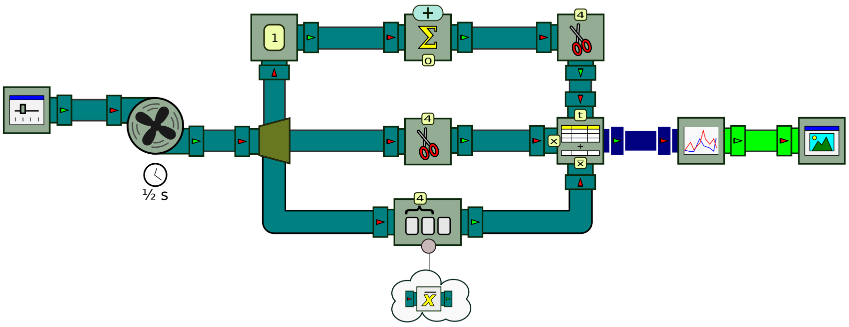


Figure 1 Example BeepBeep chain.

3.8 Probabilistic and Predictive Stream Reasoning

Fredrik Heintz (Linköping University, SE)

This talk describes stream reasoning in ProbSTL which is an extension of Signal Temporal Logic (STL) to deal with stochastic signals and predictions. This allows the logic to make probabilistic, introspective and anticipatory statements about uncertain continuous signals, which is very important for monitoring of autonomous systems operating in the physical world. The presentation is based on the paper Incremental Reasoning in Probabilistic Signal Temporal Logic by Mattias Tiger and Fredrik Heintz published in the International Journal of Approximate Reasoning 2020.

3.9 Stream Reasoning: A Tutorial

Fredrik Heintz (Linköping University, SE)

Stream reasoning is the research area that deals with the problem of performing incremental reasoning over rapidly changing information. In this tutorial we give an overview of the area including the most relevant approaches such as C-SPARQL, CQELS, EP-SPARQL, LARS, Laser, Ticker, BigSR, and MTL-based stream reasoning.

3.10 Semantic Stream Reasoning For Online Visual Sensor Fusion

Danh Le Phuoc (TU Berlin, DE)

Driven by deep neural networks (DNN), the recent development of computer vision makes visual sensors such as stereo cameras and Lidars becoming ubiquitous in autonomous cars, robotics and traffic monitoring. However, due to certain operational constraints, a processing pipeline like object tracking has to hard-wire an engineered set of DNN models to a fixed processing logic. To remedy this problem, we propose a novel semantic reasoning approach that uses stream reasoning programmes for representing commonsense and domain knowledge using non-monotonic rules in Answer Set Programming (ASP) where uncertainty of probabilistic inference operations is incorporated by weights. Our approach is realised by a dynamic reasoning framework which enables probabilistic planning to adapt the sensor fusion pipeline under operational constraints expressed in ASP. Via this talk, we will share our current implementation experience and experiment results together with our visions towards open challenges on this research direction.

3.11 Distributed Event-Based Systems: A Tutorial

Ruben Mayer (TU München, DE) and Avigdor Gal (Technion – Haifa, IL)

Distributed event-based systems offer a well-established way to gain high-level insights from low-level streaming data in real-time. These systems may come in different flavors and stem from different communities, yet they share common principles and concepts. In this tutorial, we overview these common concepts. In addition, we focus on the concept of windowing, the notion of time and the trade-off between latency and accuracy.

3.12 Streaming Graph Partitioning

Ruben Mayer (TU München, DE)

Joint work of Ruben Mayer, Kamil Orujzade, Hans-Arno Jacobsen
Main reference Ruben Mayer, Kamil Orujzade, Hans-Arno Jacobsen: "2PS: High-Quality Edge Partitioning with Two-Phase Streaming", CoRR, Vol. abs/2001.07086, 2020.
URL https://arxiv.org/abs/2001.07086

Graph partitioning is an important preprocessing step to distributed graph processing. In edge partitioning, the edge set of a given graph is split into k equally-sized partitions, such that the replication of vertices across partitions is minimized. Streaming is a viable approach to partition graphs that exceed the memory capacities of a single server. The graph is ingested as a stream of edges, and one edge at a time is immediately and irrevocably assigned to a partition based on a scoring function. However, streaming partitioning suffers from the uninformed assignment problem: At the time of partitioning early edges in the stream, there is no information available about the rest of the edges. As a consequence, edge assignments are often driven by balancing considerations, and the achieved replication factor is comparably high. In this paper, we propose 2PS, a novel two-phase streaming algorithm for high-quality edge partitioning. In the first phase, vertices are separated into clusters by a lightweight streaming clustering algorithm. In the second phase, the graph is re-streamed and edge partitioning is performed while taking into account the clustering of the vertices from the first phase. Our evaluations show that 2PS can achieve a replication factor that is comparable to heavy-weight random access partitioners while inducing orders of magnitude lower memory overhead.

3.13 Interval Temporal Logic

Angelo Montanari (University of Udine, IT)

Joint work of Laura Bozzelli, Adriano Peron, Pietro Sala, and many others

In the talk, I give a gentle introduction to interval temporal logic. I start with a short account of its distinctive features as well as of interval modalities. Then, I present the general picture of the satisfiability and model checking problems for interval temporal logic. Links to more detailed presentations are provided for interested people. Next, I briefly compare the expressiveness of interval temporal logic (in model checking) with that of LTL, CTL, and CTL*, and describe a generalization of the proposed model checking framework with regular expressions. In the last part of the talk, I outline recent and ongoing research work. In particular, I introduce and briefly compare interval temporal logics of prefixes, suffixes, and infixes, suggest possible ways of going beyond finite Kripke structures in model checking, and illustrate the problem of model checking a single interval model. I conclude the talk by discussing the appropriateness of interval temporal logic for composite event recognition.

3.14 Modular Materialisation and Incremental Reasoning

Boris Motik (University of Oxford, GB)

Maintenance of materialisation of Datalog programs plays a key role in many applications of stream reasoning. In our recent work in the KRR group at Oxford University, we have developed and thoroughly evaluated a number of algorithms that can efficiently address this problem on a range Datalog programs commonly found in practice. Despite this progress, certain programs can still be hard for incremental materialisation; for example, programs containing rules that axiomatise a binary predicate as transitive can be hard.

In this talk, I will present an outline of some of our recent work on modular materialisation and incremental maintenance of Datalog programs. We combine standard seminaive Datalog evaluation with custom modules that implement the semantics of a program subset in an arbitrary (presumably more efficient) way. We thus obtain a general framework that can integrate specialised algorithms (e.g., algorithms for the maintenance of transitive closure) with general Datalog reasoning. I will present the results of a performance evaluation showing the benefits of such a hybrid approach.

3.15 Trade-offs in Static and Dynamic Evaluation of Hierarchical Queries

Dan Olteanu (University of Oxford, GB)

Joint work of Dan Olteanu, Ahmet Kara, Milos Nikolic, Haozhe Zhang

In this talk I will discuss trade-offs in static and dynamic evaluation of hierarchical queries with arbitrary free variables. In the static setting, the trade-off is between the time to partially compute the query result and the delay needed to enumerate its tuples. In the dynamic setting, I also consider the time needed to update the query result in the presence of single-tuple inserts and deletes to the input database.

I put forward one evaluation approach that unifies both settings. This approach observes the degree of values in the database and uses different computation and maintenance strategies for high-degree and low-degree values. For the latter it partially computes the result, while for the former it computes enough information to allow for on-the-fly enumeration.

The main result of this work defines the preprocessing time, the update time, and the enumeration delay as functions of the light/heavy threshold and of the factorization width of the hierarchical query. By conveniently choosing this threshold, the approach can recover a number of prior results when restricted to hierarchical queries.

3.16 Instance Trees: A data structure for complex logical expressions

Thomas Prokosch (LMU München, DE)

Complex-event recognition relies upon storing and retrieving complex expressions (representing events) in a time- and space-efficient manner. This presentation introduces ongoing research on such a data structure: Instance Trees. The presentation first motivates, then sketches the concepts of this data structure.

3.17 Elevating the Edge to be a Peer of the Cloud

Umakishore Ramachandran (Georgia Institute of Technology – Atlanta, US)

Joint work of Umakishore Ramachandran, Harshit Gupta, Adam Hall, Enrique Saurez, Zhuangdi Xu
Main reference Umakishore Ramachandran, Harshit Gupta, Adam Hall, Enrique Saurez, Zhuangdi Xu: "Elevating the Edge to Be a Peer of the Cloud", in Proc. of the 12th IEEE International Conference on Cloud Computing, CLOUD 2019, Milan, Italy, July 8-13, 2019, pp. 17–24, IEEE, 2019.
URL https://doi.org/10.1109/CLOUD.2019.00016

Technological forces and novel applications are the drivers that move the needle in systems and networking research, both of which have reached an inflection point. On the technology side, there is a proliferation of sensors in the spaces in which humans live that become more intelligent with each new generation. This opens immense possibilities to harness the potential of inherently distributed multimodal networked sensor platforms (aka Internet of Things – IoT platforms) for societal benefits. On the application side, large-scale situation awareness applications (spanning healthcare, transportation, disaster recovery, and the like) are envisioned to utilize these platforms to convert sensed information into actionable knowledge. The sensors produce data 24/7. Sending such streams to the cloud for processing is sub-optimal for several reasons. First, often there may not be any actionable knowledge in the data streams (e.g., no action in front of a camera), wasting limited backhaul bandwidth to the core network. Second, there is usually a tight bound on latency between sensing and actuation to ensure timely response for situation awareness. Lastly, there may be other non-technical reasons, including sensitivity for the collected data leaving the locale. Sensor sources themselves are increasingly becoming mobile (e.g., self-driving cars). This suggests that provisioning application components that process sensor streams cannot be statically determined but may have to occur dynamically.

All the above reasons suggest that processing should take place in a geo-distributed manner near the sensors. Fog/Edge computing envisions extending the utility computing model of the cloud to the edge of the network. We go further and assert that the edge should become a peer of the cloud. This talk is aimed at identifying the challenges in accomplishing the seamless integration of the edge with the cloud as peers. Specifically, we want to raise questions pertaining to (a) frameworks (NOSQL databases, pub/sub systems, distributed programming idioms) for facilitating the composition of complex latency sensitive applications at the edge; (b) geo-distributed data replication and consistency models commensurate with network heterogeneity while being resilient to coordinated power failures; and (c) support for rapid dynamic deployment of application components, multi-tenancy, and elasticity while recognizing that both computational, networking, and storage resources are limited at the edge.

References

1 Zhuangdi Xu, Sayan Sinha, Harshil Shah, and Umakishore Ramachandran. *Space-Time Vehicle Tracking at the Edge of the Network*. ACM Mobicom Workshop on Hot Topics in Video Analytics and Intelligent Edges (HotEdgeVideo'19), October 2019, Los Cabos, Mexico.

2 Umakishore Ramachandran, Harshit Gupta, Adam Hall, Enrique Saurez Apuy, and Zhuangdi Xu. *Elevating the Edge to be a Peer of the Cloud*. IEEE International Conference on Cloud Computing (CLOUD 2019), July 9-12, 2019, Milano, Italy.

3 Adam Hall and Umakishore Ramachandran. *An Execution Model for Serverless Functions at the Edge*. ACM/IEEE IoT Design and Implementation (IoTDI), April 16-18, 2019, Montreal, Canada.

4 Harshit Gupta, Zhuangdi Xu, and Umakishore Ramachandran. *DataFog: Towards a Holistic Data Management Platform for the IoT Age at the Network Edge*. USENIX Workshop on Hot Topics in Edge Computing, HotEdge 2018, , Boston, MA, July 10, 2018.

5 Harshit Gupta and Umakishore Ramachandran. *FogStore: A Geo-Distributed Key-Value Store Guaranteeing Low Latency for Strongly Consistent Access*. Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems (DEBS '18) , June 2018, Hamilton, New Zealand.

6 Zhuangdi Xu, Harshit Gupta, and Umakishore Ramachandran. *STTR: A System for Tracking All Vehicles All the Time At the Edge of the Network*. Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems (DEBS '18), June 2018, Hamilton, New Zealand.

7 Enrique Saurez, Kirak Hong, Dave Lillethun, Beate Ottenwaelder, Umakishore Ramachandran. *Incremental Deployment and Migration of Geo-Distributed Situation Awareness Applications in the Fog*. ACM DEBS '16, June 20 – 24, 2016, Irvine, CA, USA. Note: winner of the "Honorable mention" award at DEBS 2016.

3.18 Time-sensitive Complex Event Processing

Kurt Rothermel (Universität Stuttgart, DE)

For many CEP applications, a limited end-to-end latency is of paramount importance. Buffering of events in front of operators is a major source of end-to-end latency. An increasing load may cause buffers to fill up rapidly leading to higher end-to-end latencies.

One approach to reduce latency is to allocate more compute resources for the operators in the CEP network. A promising way to do that is data parallelism, i.e. the input streams of an operator are partitioned and each partition is executed by a single operator instance in parallel with the other partitions. The challenges associated with this approach are manifold, such as appropriate models to predict overload, to decide where to allocate additional resources in the network of operators, or to determine how much resources should be added. Similar problem arises when the load is decreasing and resources can be deallocated.

Another approach to reduce the end-to-end latency is load shedding. This is the only alternative if resources are scarce (e.g., mobile devices or fog) or the monetary budged limits the available compute resources. The goal is to shed no more than needed to meet the given latency bound. Moreover, shedding should be done in a way so that the quality of CEP processing suffers least. Load shedding in CEP is associated with various interesting questions related to "Where to shed, how much to shed and what to shed" (e.g., which events or partial matches).

In the talk, we will report about some of our research results and ongoing work in this field.

3.19 Applications & Requirements of CER: A Tutorial

Sabri Skhiri (EURA NOVA – Mont-Saint-Guibert, BE)

The CER/CEP have been on the market for more than 10 years. However, we have seen the last five years the emergence of a new class of real time use cases. In these use cases, we have seen a significant increase of the event throughput, a need for expressing new queries, and a need to make these CER usable and manageable in operations.

The aim of this tutorial is to give an overview of this new generation of use cases while answering these questions: (1) Which are the application domains of CER? (2) What are the key requirements of CER concerning data models, recognition language expressiveness, performance (latency, throughput, predictive accuracy)? (3) How do existing approaches address these requirements? (4) What are the classes of applications that can take advantage of CER? We answer these questions by first describing the typical Streaming architecture where CER are deployed. Then, we illustrate these challenges within Industrial use cases in crowd management, banking, Telecom, Security & Surveillance and finally SOA & microservice architecture.From these cases, we summarise the key requirements and the opportunities for contributions in CER/CEP. Finally, in order to discuss the open challenges in research, we start from the open challenges in Stream processing and we project them on CER/CEP. Interestingly, the same challenges apply but in a completely different manner.

3.20 CER in Data Management: A Tutorial

Martin Ugarte (Millenium Institute – Santiago de Chile, CL) and Cristian Riveros (PUC – Santiago de Chile, CL)

Joint work of Grez, Alejandro; Vansummeren, Stijn
Main reference Alejandro Grez, Cristian Riveros, Martín Ugarte: "A Formal Framework for Complex Event Processing", in Proc. of the 22nd International Conference on Database Theory, ICDT 2019, March 26-28, 2019, Lisbon, Portugal, LIPIcs, Vol. 127, pp. 5:1–5:18, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019.
URL http://dx.doi.org/10.4230/LIPIcs.ICDT.2019.5
Main reference Martín Ugarte, Stijn Vansummeren: "On the Difference between Complex Event Processing and Dynamic Query Evaluation", in Proc. of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management, Cali, Colombia, May 21-25, 2018, CEUR Workshop Proceedings, Vol. 2100, CEUR-WS.org, 2018.
URL http://ceur-ws.org/Vol-2100/paper13.pdf
Main reference Muhammad Idris, Martín Ugarte, Stijn Vansummeren, Hannes Voigt, Wolfgang Lehner: "Efficient Query Processing for Dynamically Changing Datasets", SIGMOD Rec., Vol. 48(1), pp. 33–40, 2019.
URL https://doi.org/10.1145/3371316.3371325

Composite Event Recognition (CER) has emerged as the unifying field for technologies that require processing and correlating distributed data sources in real-time. CER finds applications in diverse domains, which has resulted in a large number of proposals for expressing and processing complex events. In this context, the objective of the tutorial is to give a theoretical perspective of the most common features found in CER. We will start by presenting a basic setting for CER that will serve to discuss what are the fundamental properties that, from a Data-Management perspective, could be asked from a CER language: well-defined sytax and semantics, composability, and denotational declarative semantics. These properties will then be exemplified by means of a particular language called CEL, which will also be used to present the main operations found in CER systems. We will discuss what are the challenges associated to defining these operations formally while satisfying the mentioned properties. Having a principled perspective on CER languages, we will move to the problem of evaluating these languages. We will discuss what are the relevant notions of efficiency and complexity, to then present the kind of lower bounds that can be obtained for evaluating CER patterns. Finally, we will show particular examples that will serve to introduce fundamental open problems.

3.21 Complex Event Recognition in Business Process Management: A Tutorial

Matthias Weidlich (HU Berlin, DE)

Business processes represent consumers as well as producers of events in many application scenarios. Common process modelling languages, therefore, include constructs to incorporate events. At the same time, event-based systems may be used as a basis for process execution and the analysis of processes. Against this background, the tutorial reviews the relation between the fields of business process management and complex event recognition. In particular, opportunities for research at the intersection of the two fields are outlined.

3.22 Short Introduction to Dynamic Complexity Theory

Thomas Zeume (TU Dortmund, DE)

Joint work of Samir Datta, Anish Mukherjee, Nils Vortmeier, Thomas Schwentick
Main reference Samir Datta, Anish Mukherjee, Nils Vortmeier, Thomas Zeume: "Reachability and Distances under Multiple Changes", in Proc. of the 45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic, LIPIcs, Vol. 107, pp. 120:1–120:14, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018.
URL http://dx.doi.org/10.4230/LIPIcs.ICALP.2018.120
Main reference Samir Datta, Raghav Kulkarni, Anish Mukherjee, Thomas Schwentick, Thomas Zeume: "Reachability Is in DynFO", J. ACM, Vol. 65(5), pp. 33:1–33:24, 2018.
URL http://dx.doi.org/10.1145/3212685

Dynamic descriptive complexity theory studies how query results can be updated in a highly parallel fashion, that is, by constant-depth circuits or, equivalently, by first-order formulas, or by the relational algebra. After gently introducing dynamic complexity theory, I will discuss recent results regarding the dynamic complexity of the reachability query.

4 Working groups

4.1 Pattern induction and composite event forecasting

Daniele Dell'Aglio (Universität Zürich, CH)

Working group participants:

- Han van der Aa (HU Berlin, DE).
- Alexander Artikis (University of Piraeus, GR & NCSR Demokritos, GR).
- François Bry (LMU München, DE).
- Daniele Dell'Aglio (Universität Zürich, CH).
- Emanuele Della Valle (Polytechnic University of Milan, IT).
- Avigdor Gal (Technion – Haifa, IL).
- Minos Garofalakis (TU Crete, GR).
- Fredrik Heintz (Linköping University, SE).
- Annika M. Hinze (University of Waikato, NZ).
- Kurt Rothermel (Universität Stuttgart, DE).
- Matthias Weidlich (HU Berlin, DE).
- Holger Ziekow (HFU – Furtwangen, DE).

Machine learning is a powerful framework to process and analyse temporal and dynamic data and can find application in composite event recognition as well. In this session we focused on two tasks. The first is composite event specification learning: what are the rules that define relevant composite events? The second is composite event forecasting: what are the next events (or composite events) that will appear from the stream? We identified in hybrid intelligence and explainability two of the main benefits that machine learning and CER can lead when combined.

Defining composite events is a challenging task, which requires domain knowledge, as well as an in-depth understanding of the data itself. As data is increasing in size, variety and heterogeneity over time, the complexity to define the composite events is growing as well.

For example, it is often the case that almost matching composite events are not recognised, as minimal discrepancies between the event definition and the input stream are sufficient to lead to negative responses. In such a scenario, machine learning can provide useful support to domain experts. For example, a human may identify the variables of interest, using machine learning to infer correlations among them. A complementary approach consists in combining human- and machine-defined rules, where the former introduces rules related to the domain knowledge, and the latter can timely infer rules related to data shifts and drifts.

At the same time, composite events represent knowledge which can be interpreted by humans and is a potential solution to achieve explainable machine learning processes. We envision the adoption of composite event formalisms in the scenarios where the input are event streams since composite events allow expressing temporal relations.

Combining machine learning and techniques for composite event recognition also poses challenges for future research. Machine learning processes require a training phase, where they learn regularities and patters by observing the input data. We traditionally distinguish between offline and online learning, depending on the fact that the training happens before or during the analysis of the stream. This depends on the type of machine learning technique, as well as the data itself. When data shows features that do not vary over time, or that vary regularly, offline learning usually leads to better outcomes. However, when the data is characterised by shifts and drifts, the data used to learn become stale over time, degrading the quality of the analyses over time. In those cases, online learning may be a more suitable solution. Creating online learning techniques, however, is challenging, since the construction of the model may be time-consuming, too slow with regards to the drifts happening in the data, or may require the definition of complex scheduling policies to retrain the algorithms. A relevant dimension to consider when thinking about machine learning and composite event recognition is distribution. In scenarios where data has high throughput, is physically distributed and controlled by different stakeholders, distributed techniques may bring several advantages. Relevant phenomena could be identified on the edge, exploiting, for example, federated learning and function shipping solutions to push decentralisation. This may also lead advantages in terms of privacy, robustness and efficiency. However, the distribution requires coordination and reconciliation mechanisms, leading to overheads and not applicable to every use case.

A final challenge which can stimulate future research is the problem of monitoring. In scenarios like security, smart cities and fraud detection, monitoring the stream leads to changes in the stream itself. For example, knowing the situation of the city during traffic jams can lead drivers to opt for different paths, with the potential risk of generating new traffic jams. Possible ways to tackle this challenge are the design of cost functions that focus on the community, e.g. a navigation system may try to optimize the average speed of all the cars in the city. Another solution may come from game theory through the identification of equilibrium points. Another problem related to monitoring is the evaluation: changes in the input stream make it hard to test and systematically compare alternative solutions. One could follow analytical analyses, or exploit simulators to run experimental tests. However, both approaches may be complex and costly to follow.

4.2 Process strategies, parallelization and geo-distribution

Daniele Dell'Aglio (Universität Zürich, CH)

Working group participants:
- Daniele Dell'Aglio (Universität Zürich, CH).
- David Eyers (University of Otago, NZ).
- Minos Garofalakis (TU Crete, GR).
- Manfred Hauswirth (Fraunhofer FOKUS – Berlin, DE).
- Alessandro Margara (Polytechnic University of Milan, IT).
- Ruben Mayer (TU München, DE).
- Umakishore Ramachandran (Georgia Institute of Technology – Atlanta, US).
- Till Rohrmann (Ververica – Berlin, DE).
- Kurt Rothermel (Universität Stuttgart, DE).
- Sabri Skhiri (EURA NOVA – Mont-Saint-Guibert, BE).
- Riccardo Tommasini (University of Tartu, EE).

Nowadays, networks are usually complex environment, with nodes heterogeneous in computational power, storage capabilities, network connections, geographical locations and ownership. This is the case of modern edge infrastructures, where networks have powerful central nodes (private or public cloud infrastructures) and nodes with limited resources at the edge.

Moreover, network failures (e.g. congestion and broken nodes) bring dynamics in such networks. Dynamics also happen in specific contexts, such as in automotive, where edge nodes move and affect the network topology, varying the connections among the network nodes. On top of those networks, stakeholders want to detect events of interests, based on either the data collected by a node itself or the data exchanged by other nodes.

To recognize composite events over these heterogeneous and dynamic networks, we need flexible processing frameworks that expose the following features. Such frameworks should be flexible, able to move both the data and the recognizing functions across the network. Flexibility is important for runtime performance, to improve time performance metrics (e.g. latency and throughput), as well for privacy purposes (e.g. process the data locally). Another feature processing frameworks should offer robustness. The frameworks should be able to cope with network failures, to do not stop the execution if a node breaks or connection is congested, by exploiting, for example, redundancy. At the same time, the processing should be robust to the presence of noise and wrong data. Detecting the same composite events with data observed by different nodes may offer different perspectives, allowing to detect and isolate wrong results. Finally, such frameworks should be observable, to monitor the execution of the system and potentially detect failures and issues.

To design and build such processing frameworks, we should take into account the existence of features that can hardly co-exist in the same solution. There exists a trade-off between expressiveness and performance: the more expressive the constructs in the programming frameworks, the more complex the business logic, with subsequent loss of performance. There is also a trade-off between generalization and specialization: domain-specific solutions may exploit the context to introduce specific optimizations and solutions, at the price of being usable only in a specific number of use cases. The framework should be able to detect composite events in a continuous fashion, coping with issues in the input streams, such as noise and out of orders. The detection should also be distributed, with the need for consensus mechanisms to reach agreements on the identification of the events. It is essential to observe that different nodes may have different perceptions of reality, which could be captured through local ontologies.

4.3 Expressiveness, Compositionality & Hierarchies, and Common Framework

Boris Motik (University of Oxford, GB) and Martin Ugarte (Millenium Institute – Santiago de Chile, CL)

Working group participants:

- Alexander Artikis (University of Piraeus, GR & NCSR Demokritos, GR).
- François Bry (LMU München, DE).
- Emanuele Della Valle (Polytechnic University of Milan, IT).
- Thomas Eiter (TU Wien, AT).
- Alessandro Margara (Polytechnic University of Milan, IT).
- Angelo Montanari (University of Udine, IT).
- Boris Motik (University of Oxford, GB).
- Thomas Prokosch (LMU München, DE).
- Tore Risch (Uppsala University, SE).
- Cristian Riveros (PUC – Santiago de Chile, CL).
- Kostas Stathis (Royal Holloway, University of London, GB).
- Martin Ugarte (Millenium Institute – Santiago de Chile, CL).
- Stijn Vansummeren (Université Libre de Bruxelles, BE).
- Matthias Weidlich (HU Berlin, DE).
- Thomas Zeume (TU Dortmund, DE).

This is a summary of the outcomes of two discussion sessions held during the Dagstuhl 20071 seminar on Foundations of Composite Event Recognition. The topics of the two sessions were quite related, so it is natural to summarise them jointly.

4.3.1 Expressiveness, Compositionality & Hierarchies

The discussion in this session was motivated by an observation that the field of Complex Event Recognition (CER) is very broad and diverse, which makes understanding the relationships between various approaches proposed in the literature quite difficult. The discussion revolved around a number of issues, as summarised next.

CER from an abstract perspective. It was observed that the community has not agreed on a common abstract perspective of the CER problem. The following three possibilities have been proposed and discussed.

- CER is a *model checking problem* – that is, the problem of verifying whether a finite input satisfies a particular property. After a discussion, there was consensus that this perspective most likely does not adequately capture CER.

- CER is a *monitoring problem* – that is, the problem of detecting the instant when monotonically increasing input satisfies a particular property. There was a sentiment that a minority of CER applications may fall into this category.
- CER is a *synthesis problem* – that is, the problem of transforing an infinite input into an infinite output using a predetermined specification. This view was identified as closest to most CER applications.

Providing a common model. It was noted that agreeing on a common model had a tremendous impact in areas such as databases and description logics, and so doing the same in the context of CER might produce similar benefits: it would allow for an easier comparison of the expressiveness and the capabilities of different approaches, both formally and informally, and standardisation might foster interoperability of tools and systems produced by different groups. To achieve these goals, both data and query models should be agreed upon. This raised a question of what should CER systems produce as output, and the following views were put forward.

- An answer is a *sequence of time-annotated facts.*
- An answer is a sequence of *time-annotated sets of tuples.* That is, database queries produce sets of tuples, so by analogy CER systems should produce streams of sets of tuples.
- An answer is *a sequence of time-annotated sets of facts from the input* that constitute a complex event.

Who would use a common model? It was noted that defining a common model was difficult partly because of a lack of clarity about who its intended users would be. The following possibilities were discussed.

- A common model would be used by end-users (i.e., practitioners) in the field. In this view, a common model would play the role analogous to SQL by defining an interface that CER systems would provide. In such a case, an important design guideline would be to produce an intuitive model that closely reflects the end-users' problems.
- A common model would be used mainly by researchers as an agreed-upon yardstick for analysing various approaches from a conceptual and/or theoretical standpoint. Thus, a common model would have to be very expressive and general, whereas its practical applicability would be less important. It was noted that such a model would play a role similar to that of a Turing machine.
- A common model would be used by both end-users and researchers. This would be an ideal outcome, but it might be difficult to attain.

The role of time. There was considerable discussion about the role of time in such a model. The following opinions were voiced.

- Time is *not special.* In this view, time is just another piece of data that may or may not be present in the stream elements. The common model could be just the relational model, where time instants and intervals could be modelled using one or two temporal attributes, respectively.
- Time is *immanent* to CER, and it provides and *orthogonal dimension* to the stream content. That is, the data in the stream can be seen as 'opaque': we just need to be able to manipulate data items using an appropriate algebra. For example, data items can be relational, in which case they are manipulated using the relational algebra; or data items can be XML documents, in which case they are manipulated using XPath and XQuery. A CER system can be built on top of any data model by considering a time-annotated sequence of data items. The CER system should provide constructs for manipulating the

temporal component of the stream, which can be integrated with the underlying algebra in a modular fashion. It was observed that this underpins the CQL approach by Widom et al.

Time instants vs. time intervals. There was a long discussion about whether a common model for CER should be based on time instants or time intervals. The following arguments were put forward in favour of each view.

- In the former view, input events are instantaneous occurrences on a discrete, partially ordered timeline. Complex events can be thought of as patterns in the input, and they are also instantaneous in the sense that they occur at instants at which the patterns are recognized. In this way, there is no distinction between input events and complex events, so the latter can be used as input to create event hierarchies. It was argued that such a viewpoint is appropriate for CER because sensor readings are instantaneous and the observed value between two successive readings is unknown. Finally, it was argued that duration of an event can be defined as the time period between the event's onset and cessation. For example, one can identify the time points at which the 'Fire is detected' event commences and stops, which gives the duration of the event.
- The latter view assumes that events are inherently durative in nature, so a CER system should have the notion of a duration built into it. It was pointed out that this view is more general than the time point view, and it was argued that durations are strictly necessary for temporal aggregation. Finally, it was argued that an interval view might provide a more natural user interface to a CER system.

Proposals for a common model. Several different formalisms were proposed as candidates for a common model.

- Relational model seemed to be a natural candidate. It was argued that no specific treatment of time would be needed in such a setting as temporal information could be encoded using one (for point-based events) or two (for interval-based events) temporal attributes.
- First-order logic could be used analogously, and it could be extended with second-order features such as Kleene closure.
- Various temporal logics (either point- or interval-based) were also put forward as natural candidates.

A key question was what kind of object should be produced by query evaluation. One view was that queries can be seen as formulas with free variables, so answers are variable substitutions that make the formula true.

A common model is not needed. It was also argued that agreeing on a common model for CER might be too difficult or even undesirable. The argument was that the model should be chosen to fit the application at hand; however, the application space seems to be too heterogenous to be unified, so it might be difficult to come up with a one-size-fits-all solution. Instead, different approaches should be compared directly – for example, by providing pairwise translations between approaches.

4.3.2 Common Framework

As the discussion in the 'Expressiveness, Compositionality, & Hierarchies' session revealed, reaching agreement in the community on a single common model for CER might be too ambitious at this point. Therefore, the discussion in the 'Common Framework' session explored the possibility of providing a common *meta-model* for CER. The objective was to present a more abstract view of CER that would clearly identify key elements of most CER systems. Specifically, this meta-model would specify what a CER system *does*, without focussing on *how* this is done. The meta-model would not necessarily focus on a specific language for encoding events and queries; rather, it would provide a list of conceptual components that could be instantiated in different ways. The main benefit of such a meta-model would be to provide a common way of thinking and talking about CER, which would make discussions in the community easier.

The rest of this section summarises the meta-model that was proposed and discussed at the seminar.

4.3.2.1 Abstract View of CER

Seen from an abstract point of view, a CER system observes an ever-increasing amount of information about an application environment. This information is delivered to the CER system in discrete, finite chunks called *updates*. Thus, at each instant i, a CER system has observed a finite number of such updates. The task of a CER system is to extract useful information from all information observed up to a given instant. More precisely, a CER system should exhibit behaviour that looks as if the following steps were performed at each instant i.

1. All observed updates are *combined* into a *world view* W_i. The role of W_i is to reflect the history (or, in some cases, the relevant part of the history) of the application environment. Often, this step will involve *background knowledge* K about the environment that, for all intents and purposes, can be assumed to be immutable.
2. A fixed *query* Q is evaluated over W_i, and the answer A_i is sent to the user at time instant i. The job of Q is to specify which part of the world view is relevant for the application at hand. In other words, Q is a function that extracts useful information from W_i.

This idea can be summarised formally as follows. A CER system is parameterised by immutable background knowledge K, an aggregation function AGG, and a query Q. Function AGG must be applicable to K and a finite sequence of *updates* $U_1.U_2 \ldots U_i$, and query Q is a function that must be applicable to the aggregation result. Then, given an infinite *stream* $S = U_1.U_2 \ldots$ of updates, the job of a CER system is to produce the infinite sequence of answers $A_1.A_2 \ldots$ where, for each $i \geq 1$, we have $A_i = Q(W_i)$ for $W_i = \mathsf{AGG}(K, U_1.U_2 \ldots U_i)$.

It is crucial to understand that this specification *does not* mandate that the system necessarily has to compute W_i and then evaluate Q on W_i at each instant i. Rather, this specification only specifies what the observable behaviour of the system should be, and the system is free to choose any appropriate implementation/evaluation strategy. In fact, it is usually reasonable to introduce various assumptions about the properties of the computation that a CER system should use. We call such assumptions *nonfunctional requirements* in order to stress that these specify certain aspects of a system's operation, rather than restrict the answers that the system must compute. For example, a common nonfunctional requirement might be that a CER system should use a bounded amount of memory during its operation. Such requirements will determine whether a CER system can correctly process the query Q or not.

It is useful to further assume that background knowledge, updates, and queries are all expressed in appropriate *knowledge*, *update*, and *query languages* $\mathcal{L}_K$, $\mathcal{L}_U$, and $\mathcal{L}_Q$, respectively. Formally, these languages can be seen as classes whose members constitute legal values for K, U_i, and Q. In some cases it might be useful to restrict not just the form of each update, but also to place certain *structural constraints* on the stream itself. For example, a structural constraint might be ‘The time stamps cannot be decreasing’. Then, the objective of CER research can be framed as the task of studying the algorithmic methods that produce the correct answers for specific combinations of nonfunctional requirements, languages $\mathcal{L}_K$, $\mathcal{L}_U$, and $\mathcal{L}_Q$, and structural constraints on the stream.

4.3.2.2 Examples

To clarify these ideas, this section presents several very simple instantiations of the CER meta-model.

▶ **Example 1.** Let the update language $\mathcal{L}_U$ consist of all finite sets of relational facts of the form $R(a_1, \ldots, a_n, t)$, where R is a relation, all a_j are constants, and t is a time point. In other words, each update U_i is a finite set of relational facts with a time stamp. Moreover, ignoring any question of background knowledge for the moment, let the aggregation function be defined by

$$\mathsf{AGG}(U_1.U_2 \ldots U_i) = \{Now(i)\} \cup \bigcup_{j=1}^{i} U_j. \tag{1}$$

In other words, AGG takes the union of all updates, but it also introduces a fact that represents the current time point. Each world view W_i thus contains all information that the CER system observed up to instant i, as well as information about where in the stream we are.

As an example, consider the stream that consists of the following updates:

$$U_1 = \{temp(burner_1, 40, 1)\} \tag{2}$$

$$U_2 = \{temp(burner_2, 20, 2)\} \tag{3}$$

$$U_3 = \{temp(burner_1, 60, 2),\ temp(burner_2, 45, 3)\} \tag{4}$$

Intuitively, each update represents temperature readings of gas burners, where each reading is associated with an instant at which the reading has been produced. Note that the instants inside the updates do not necessarily correspond to the instants at which an update has been received. For example, update U_3 is received at instant 3, but fact $temp(burner_1, 60, 2)$ refers to time instant 2. We next investigate languages for querying W_i.

First-order logic provides the formal foundations for a substantial part of SQL, so it is reasonable to try to use it in a streaming setting. Thus, let us define queries over W_i as domain-independent first-order formulas with free variables; moreover, we define an answer to a query on W_i as the set of all substitutions of the free variables that make the query true on W_i. Such a language is very expressive, and in particular it allows us to ask questions about both past and present instants. For example, the following query identifies the most recent time instant after which the temperature reading for $burner_1$ was above 35 for two consecutive instants:

$$\begin{aligned}
&Q(t) = \exists t_{now} \exists x_1 \exists x_2 .[\\
&\quad Now(t_{now}) && \wedge \\
&\quad temp(burner_1, x_1, t) \ \wedge\ x_1 \geq 35 && \wedge \\
&\quad temp(burner_1, x_2, t+1) \ \wedge\ x_1 \geq 35 && \wedge \\
&\quad \forall t' \forall x' . (t + 2 \leq t' \leq t_{now} \wedge temp(burner_1, x', t')) \Rightarrow x' < 35 \\
&]
\end{aligned} \tag{5}$$

Note that Q can return answers about past time instants. For example, $A_3 = \{t \mapsto 1\}$ – that is, evaluating Q at time instant 3 produces an answer that refers to time instant 1. This illustrates the benefit of distinguishing time instants at which a query is evaluated from time instants that the query talks about. The former time instant determines what updates have been observed and thus plays a prominent role in the conceptual view of CER. In contrast, associating updates with time stamps can be viewed as a question of modelling – that is, they can be handled as part of the CER "model content".

▶ **Example 2.** While the query language from Example 1 is very expressive, it has a significant drawback: answering any first-order query correctly over an arbitrary stream requires storing all observed information. Therefore, it might be interesting to identify query languages for which each query can be processed on an arbitrary stream using a finite amount of memory.

A very simple way to achieve this is to evaluate each query inside a *temporal window*. For the purpose of this example, let us extend the notion of a query to a pair (φ, n), where φ is a domain-independent first-order formula, and n is an integer specifying the window aperture. To evaluate such a query over a world view W_i, one evaluates φ over the subset of all facts in W_i whose time stamp is between $t_{now} - -n$ and t_{now}, for $Now(t_{now}) \in W_i$. One can now investigate the nonfunctional requirements and the structural constraints on the stream that will allow a CER system to answering an arbitrary such query over an arbitrary stream.

Answering queries in this query language is still difficult, for at least the following two reasons.

- Updates are not bounded in size. That is, we cannot answer any first-order query using a finite amount of memory if the number of facts in each update can exceed the available memory.
- We have not placed any restriction on the relation between the instant i of an update U_i and the time stamps of the facts contained inside U_i. As a result, U_i can refer to instants that will become relevant arbitrarily far in the future, and storing all such instants may require an unbounded amount of memory.

To overcome these difficulties, we may place the following two structural constraints on the stream.

- We assume that each update U_i contains no more than ℓ facts for some fixed integer ℓ. For example, this constraint may hold in a setting where the number of sensors is limited to ℓ.
- We assume that the time stamp of each fact in U_i must be between $i - -\wp$ and i for some fixed integer $\wp$. This constraint may hold in a setting where a global clock ensures that no sensor produces "future" time stamps, and that all sensor readings are delivered to the CER system within $\wp$ time instants.

The above structural constraints on the stream may not hold in every setting. If, however, they hold, then a CER system can answer every query with window n by storing only the last n updates, which requires a bounded amount of memory. In turn, this conceptual observation opens to door to further investigation of the algorithms and data structures necessary to evaluate such queries efficiently.

▶ **Example 3.** In Examples 1 and 2, new observations were simply appended to the world view (while updating the current time stamp). Our meta-model for CER, however, allows us to also capture a setting where updates can retract information from the world view. For example, let us assume that we equip the gas burners from Example 1 with an additional sensor that can determine that a reading produced at an earlier time instant was invalid. To incorporate this, we can extend the update language so that each update is of the form

$U_i = \circ_i F_i$, where $\circ_i$ is either $+$ or $-$, and F_i is a set of time-stamped facts as in Example 1. Then, we can define the aggregation function inductively as follows, where ϵ is the empty sequence of updates:

$$\mathsf{AGG}(\epsilon) = \emptyset \tag{6}$$

$$\mathsf{AGG}(U_1.U_2 \dots U_1.U_{i-1}. + F_n) = \mathsf{AGG}(U_1.U_2 \dots U_1.U_{i-1}) \cup F_n \tag{7}$$

$$\mathsf{AGG}(U_1.U_2 \dots U_1.U_{i-1}. - F_n) = \mathsf{AGG}(U_1.U_2 \dots U_1.U_{i-1}) \setminus F_n \tag{8}$$

In other words, aggregating updates now involves both adding and retracting facts. Note that each W_i reflects the information that is believed to be true at instant i. For example, consider the following updates:

$$U_1 = +\{\mathit{temp}(\mathit{burner}_1, 40, 1)\} \tag{9}$$

$$U_2 = +\{\mathit{temp}(\mathit{burner}_2, 20, 2)\} \tag{10}$$

$$U_3 = -\{\mathit{temp}(\mathit{burner}_1, 40, 2)\} \tag{11}$$

World view W_2 contains both $\mathit{temp}(\mathit{burner}_1, 40, 1)$ and $\mathit{temp}(\mathit{burner}_2, 20, 2)$, whereas world view W_3 contains only the latter fact and thus takes into account that the former fact was found to be in error.

▶ **Example 4.** We can further extend Example 3 and assume that each update involves addition or retraction of a logical formula expressed in a temporal logic. The aggregation function can combine all updates into the current world view using belief revision operators. Finally, a query can involve checking entailment of facts/formulas from world views.

▶ **Example 5.** All examples thus far considered updates and queries encoded using symbolic languages, but the proposed CER meta-model makes no such assumptions. For example, the query can be given as a neural network, updates can consist of readings for various inputs to the network, and the aggregation function should determine how to combine different updates into the input to the network. Then, we can analyse which classes of neural networks can express what kinds of queries.

4.3.2.3 Next Steps

It was suggested during the seminar that the natural next step would be to prepare a survey paper that would (i) describe the meta-model in more detail, (ii) discuss how to instantiate this meta-model in order to capture the various proposals from the literature, and (iii) identify classes of languages and constraints found in the existing body of literature. It was suggested that such a publication might be produced jointly by the interested parties at one of relevant future events, such as the Stream Reasoning Workshop.

4.4 Benchmarking

Riccardo Tommasini (University of Tartu, EE)

Working group participants:

- Thomas Eiter (TU Wien, AT).
- David Eyers (University of Otago, NZ).
- Wim Martens (Universität Bayreuth, DE).
- Ruben Mayer (TU München, DE).
- Umakishore Ramachandran (Georgia Institute of Technology – Atlanta, US).
- Till Rohrmann (Ververica – Berlin, DE).
- Riccardo Tommasini (University of Tartu, EE).
- Stijn Vansummeren (Université Libre de Bruxelles, BE).

Domain-specific benchmarks aim to foster technological progress by guaranteeing a fair assessment [5]. To this extent, Gray's seminal work identifies important principles that drive the design of data system benchmarks.

- *Relevance*: a benchmark must measure the price/performance ration of systems when performing typical operations within its domain.
- *Portability*: a benchmark must be easy to implement on many heterogeneous systems and architectures.
- *Scalability*: a benchmark should apply to small and large computer systems.
- *Simplicity*: a benchmark must be understandable, otherwise it will lack credibility.

In addition, Karl Huppler pushed the benchmark guidelines event further. In "The art of building a good benchmark" [6], he identifies the following three additional principles.

- *Repeatability*: there is confidence that the benchmark can be run a second time with the same result.
- *Fairness*: all systems and/or software being compared can participate equally.
- *Verifiability*: there is confidence that the documented result is real.

Gray and Huppler stressed the economical aspects of benchmarking. Intuitively, a good benchmark should be representative yet sustainable for the community to adopt it. For a proper evaluation, it is not necessary to be compliant with all the listed principles but only to those reflecting the benchmarking purpose [6].

In the context of complex event recognition (CER) a consolidated benchmark is still missing. In the related literature, few attempts tried to identify use-cases, key performance indicators (KPI), and relevant challenges to the research community to address [8, 2, 1].

Nevertheless, performance evaluations are still not homogeneous. In absence of real-world event streams, researchers are forced to adapt analytic benchmarks like the Linear Road [1], database benchmarks like BEAST [3], or benchmarks designed for Message-Oriented Middleware [7]. Intuitively, this handcrafted approach to benchmarking limits the repeatability and reproducibility of the results. The cost of maintenance of the customized benchmarks is entirely on the individual research groups and, as a result, it is hard to guarantee long-term support.

To this extent, the working group focused on identifying a sustainable path with concrete operational steps that could lead to the design of a domain-specific benchmark for CEP that is maintainable by the community.

4.4.1 Types of benchmarks (relevance, simplicity, and fairness)

Initially, the discussion focused on identifying interesting types of benchmarks. Two main areas emerged, i.e., macro- (also known as use-case driven) and micro-benchmarks. The former focuses on evaluating systems with respect to specific workloads, typically inspired by real-world scenarios. The latter, instead, focuses on evaluating the performance of single operators.

Macro-benchmarks directly relate with the ongoing effort behind the DEBS Grand Challenges, which yearly provide interesting use-cases and workloads for the community to solve. On the other hand, micro-benchmarks relate to the definition of a common model for CER e/o a core algebra of operation that CER engines must support.

4.4.2 Lack of standards (portability)

Industrial and academic research on CER highlights the lack of shared data and query models towards a standardization. These agreements are crucial to develop and maintain a benchmarking suite for the community.

In particular, during the meeting is emerged that the identification of data formats and query languages is of paramount importance to move forwards any activity in the context of benchmarking.

4.4.3 Technical support (scalability, repeatability, and verifiability)

Last but not least, the working group has highlighted the issue related to technical supports. To this extent, similar criteria used for data publishing should be adopted. Community benchmarks should be FAIR [9], i.e.,

- *Findable*: it must be identifiable and registered in searchable resources.
- *Accessible*: it must be retrievable by their identifiers using open and standard protocols.
- *Interoperable*: it must use a formal and shared language for it representation and reference other relevant resources.
- *Reusable*: it must be described with a plurality of accurate and relevant attributes, e.g., licenses, and provenance metadata.

4.4.4 Conclusion and roadmap

In conclusion, the working group has identified two concrete steps towards the definition of a standard benchmark for CER. The two steps plan is described below by indicating, for each step, a minimum set of subtask and potential outcomes:

1. Design a replication study and literature analysis (including DEBS Grand Challenges)
   - Identify the dimensions of interest for the literature starting from the recent works [4];
   - identify the systems and experiments of interest for the replication study starting from the DEBS Grand Challenges;
   - consider as input the ongoing work on uniforming languages and models.
2. Start a working group that has the extent of creating a community benchmark.
   - Contact the Linked Data Benchmark Council
   - Invite people from DEBS Grand Challenge community.
   - Integrate this into the Stream Reasoning COST Action organized by Thomas Eiter.
   - Identify other academics and industrial groups.

References

1 A. Arasu, M. Cherniack, E. F. Galvez, D. Maier, A. Maskey, E. Ryvkina, M. Stonebraker, and R. Tibbetts. Linear road: A stream data management benchmark. In M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, and K. B. Schiefer, editors, *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB 2004, Toronto, Canada, August 31 – September 3 2004*, pages 480–491. Morgan Kaufmann, 2004. 10.1016/B978-012088469-8.50044-9. URL http://www.vldb.org/conf/2004/RS12P1.PDF.

2 P. Bizarro. Bicep – benchmarking complex event processing systems. In K. M. Chandy, O. Etzion, and R. von Ammon, editors, *Event Processing, 6.5. – 11.5.2007*, volume 07191 of *Dagstuhl Seminar Proceedings*. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2007. URL http://drops.dagstuhl.de/opus/volltexte/2007/1143.

3 A. Geppert, S. Gatziu, and K. R. Dittrich. A designer's benchmark for active database management systems: oo7 meets the BEAST. In T. K. Sellis, editor, *Rules in Database Systems, Second International Workshop, RIDS '95, Glyfada, Athens, Greece, September 25 - 27, 1995, Proceedings*, volume 985 of *Lecture Notes in Computer Science*, pages 309–326. Springer, 1995. 10.1007/3-540-60365-4_135. URL https://doi.org/10.1007/3-540-60365-4_135.

4 N. Giatrakos, E. Alevizos, A. Artikis, A. Deligiannakis, and M. N. Garofalakis. Complex event recognition in the big data era: a survey. *VLDB J.*, 29(1):313–352, 2020. 10.1007/s00778-019-00557-w. URL https://doi.org/10.1007/s00778-019-00557-w.

5 J. Gray, editor. *The Benchmark Handbook for Database and Transaction Systems (1st Edition)*. Morgan Kaufmann, 1991.

6 K. Huppler. The art of building a good benchmark. In R. O. Nambiar and M. Poess, editors, *Performance Evaluation and Benchmarking, First TPC Technology Conference, TPCTC 2009, Lyon, France, August 24-28, 2009, Revised Selected Papers*, volume 5895 of *Lecture Notes in Computer Science*, pages 18–30. Springer, 2009. 10.1007/978-3-642-10424-4_3. URL https://doi.org/10.1007/978-3-642-10424-4_3.

7 K. Sachs, S. Kounev, S. Appel, and A. P. Buchmann. Benchmarking of message-oriented middleware. In A. S. Gokhale and D. C. Schmidt, editors, *Proceedings of the Third ACM International Conference on Distributed Event-Based Systems, DEBS 2009, Nashville, Tennessee, USA, July 6-9, 2009*. ACM, 2009. 10.1145/1619258.1619313. URL https://doi.org/10.1145/1619258.1619313.

8 T. Scharrenbach, J. Urbani, A. Margara, E. D. Valle, and A. Bernstein. Seven commandments for benchmarking semantic flow processing systems. In P. Cimiano, Ó. Corcho, V. Presutti, L. Hollink, and S. Rudolph, editors, *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, volume 7882 of *Lecture Notes in Computer Science*, pages 305–319. Springer, 2013. 10.1007/978-3-642-38288-8_21. URL https://doi.org/10.1007/978-3-642-38288-8_21.

9 M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen,

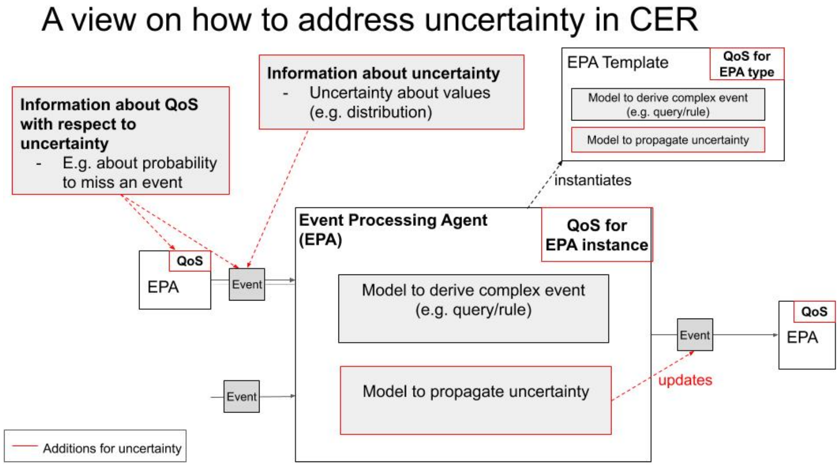

A view on how to address uncertainty in CER
Information about QoS with respect to uncertainty
- E.g. about probability to miss an event
Information about uncertainty
- Uncertainty about values (e.g. distribution)
EPA Template
QoS for EPA type
Model to derive complex event (e.g. query/rule)
Model to propagate uncertainty
instantiates
Event Processing Agent (EPA)
QoS for EPA instance
QoS
EPA
Event
Model to derive complex event (e.g. query/rule)
Model to propagate uncertainty
Event
Event
updates
QoS
EPA
Additions for uncertainty

Figure 2 Addressing uncertainty in CER

J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3 (1):160018, 2016. 10.1038/sdata.2016.18. URL https://doi.org/10.1038/sdata.2016.18.

4.5 Uncertainty in Complex Event Recognition

Han van der Aa (HU Berlin, DE), Avigdor Gal (Technion – Haifa, IL), Sylvain Hallé (University of Quebec at Chicoutimi, CA), Annika M. Hinze (University of Waikato, NZ), and Holger Ziekow (HFU – Furtwangen, DE)

Working group participants:

- Han van der Aa (HU Berlin, DE).
- Thomas Eiter (TU Wien, AT).
- Avigdor Gal (Technion – Haifa, IL).
- Sylvain Hallé (University of Quebec at Chicoutimi, CA).
- Manfred Hauswirth (Fraunhofer FOKUS – Berlin, DE).
- Fredrik Heintz (Linköping University, SE).
- Annika M. Hinze (University of Waikato, NZ).
- Danh Le Phuoc (TU Berlin, DE).
- Holger Ziekow (HFU – Furtwangen, DE).

Uncertainty is inherent in many use cases for complex event recognition (CER). This uncertainty stems from two main sources. One source is noise and errors in the input data for complex event recognition. For instance, measurements from input sensors may be incorrect or incomplete. The other source for uncertainty lies in the inference for CER. That is, the inference of a complex event from simpler events may be of probabilistic nature. For instance, a system may classify a situation as an emergency, even though it turns out to be a false alarm.

Uncertainty can also come from the deliberate intention of an event publisher to prevent complete disclosure of an event stream to its subscribers. One possible manifestation of this situation is when a source of events is subject to access control constraints, such as privacy policies. In the same way that the access to statistical databases can be restricted in order to preserve anonymity (e.g. query restriction and query perturbation [1]) by the deliberate blocking or insertion of noise, one can imagine that a data stream may be subjected to similar conditions, which themselves may depend on the past content of the stream, as in history-based access control [2]. In such cases, the reason for the presence of uncertainty is not technical or fortuitous in nature; uncertainty may even be inserted with the precise design of actively preventing some processing of the stream by downstream agents.

We argue that uncertainty cannot be avoided in many applications of CER and see the need to explicitly account for it in CER systems. In particular, we see the need for reasoning about uncertainty (of confidence) when reacting to events. If a system detects an emergency with very low confidence, we may react with different measures than if the confidence is high. However, our threshold for acting on a possible emergency is lower than for less consequential situations. CER systems should therefore provide means to make uncertainty explicit. Yet, it is common practice to simply employ – often fixed – thresholds in CER an treat detected events as if they were certain. This not only projects a false sense of certainty, if prohibits any reasoning about uncertainty when taking action. We therefore see the need to expand general frameworks for CER with means to capture and reason about uncertainty. As results, we augmented a general architecture for CER with specific extension points. These extension points show what is missing in current CER frameworks to add the dimension of uncertainty. Figure 2 provides an overview of the extended CER framework.

A range of candidate techniques for implementing the identified extension point as well as many tradeoffs amongst them exists. We believe that the CER would benefit from explicitly capturing the design options along with their strengths and weaknesses in an overarching framework.

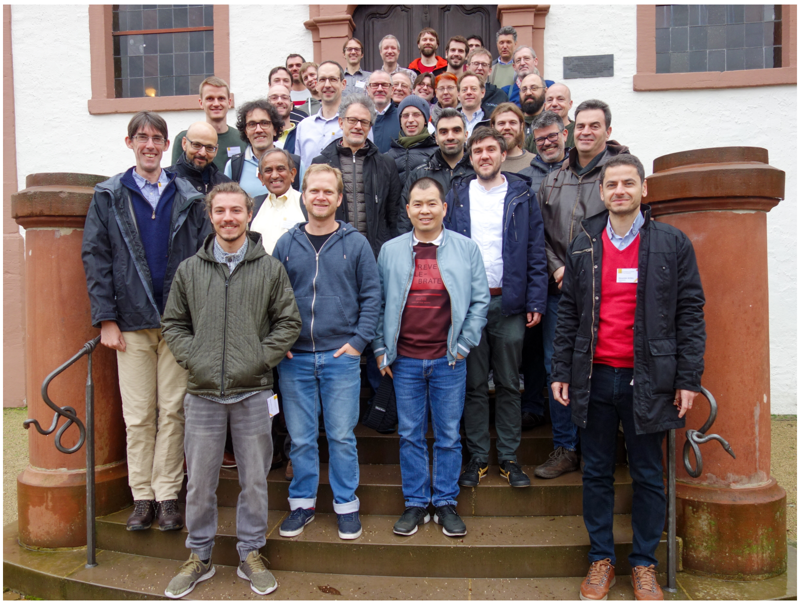
References

1 N. Adamand, J. Wortmann. *Security-control methods for statistical databases: A comparative study*. ACM Computing Surveys, 1989.

2 M. Abadi, C. Fournet. *Access control based on execution history*. In Proceedings of the 10th Annual Network and Distributed System Security Symposium, 2003.

Participants

- Elias Alevizos
  Demokritos – Athens, GR
- Alexander Artikis
  NCSR Demokritos – Athens, GR
- François Bry
  LMU München, DE
- Diego Calvanese
  Free University of
  • Bozen-Bolzano, IT
- Daniele Dell'Aglio
  Universität Zürich, CH
- Emanuele Della Valle
  Polytechnic University of Milan, IT
- Thomas Eiter
  TU Wien, AT
- David Eyers
  University of Otago, NZ
- Avigdor Gal
  Technion – Haifa, IL
- Minos Garofalakis
  Technical University of Crete – Chania, GR
- Alejandro J. Grez
  PUC – Santiago de Chile, CL
- Sylvain Hallé
  University of Quebec at Chicoutimi, CA
- Manfred Hauswirth
  Fraunhofer FOKUS – Berlin, DE
- Fredrik Heintz
  Linköping University, SE
- Annika M. Hinze
  University of Waikato, NZ
- Boris Koldehofe
  University of Groningen, NL
- Danh Le Phuoc
  TU Berlin, DE
- Alessandro Margara
  Polytechnic University of Milan, IT
- Wim Martens
  Universität Bayreuth, DE
- Ruben Mayer
  TU München, DE
- Angelo Montanari
  University of Udine, IT
- Boris Motik
  University of Oxford, GB
- Thomas Prokosch
  LMU München, DE
- Umakishore Ramachandran
  Georgia Institute of Technology – Atlanta, US
- Tore Risch
  Uppsala University, SE
- Cristian Riveros
  PUC – Santiago de Chile, CL
- Till Rohrmann
  Ververica – Berlin, DE
- Kurt Rothermel
  Universität Stuttgart, DE
- Sabri Skhiri
  EURA NOVA – Mont-Saint-Guibert, BE
- Kostas Stathis
  Royal Holloway, University of London, GB
- Riccardo Tommasini
  University of Tartu, EE
- Martin Ugarte
  Millenium Institute – Santiago de Chile, CL
- Jacopo Urbani
  VU University Amsterdam, NL
- Han van der Aa
  HU Berlin, DE
- Stijn Vansummeren
  Free University of Brussels, BE
- Matthias Weidlich
  HU Berlin, DE
- Thomas Zeume
  TU Dortmund, DE
- Holger Ziekow
  HFU – Furtwangen, DE



Report from Dagstuhl Seminar 20081

Scheduling

Edited by

Nicole Megow[1], David Shmoys[2], and Ola Svensson[3]

1 **Universität Bremen, DE,** nicole.megow@uni-bremen.de
2 **Cornell University – Ithaca, US,** david.shmoys@cornell.edu
3 **EPFL – Lausanne, CH,** ola.svensson@epfl.ch

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 20081 "Scheduling". The seminar focused on the interplay between scheduling problems and problems that arise in the management of transportation and traffic. Important aspects at the intersection of these two research directions include data-driven approaches in dynamic decision-making, scheduling in combination with routing, shared mobility, and coordination versus competition.

Seminar February 16–21, 2020 – http://www.dagstuhl.de/20081
2012 ACM Subject Classification Theory of computation → Scheduling algorithms,Applied computing → Transportation, Mathematics of computing → combinatorial optimization, Theory of computation → Approximation algorithms analysis
Keywords and phrases scheduling, optimization, approximation algorithms, routing, transportation, mechanism design
Digital Object Identifier 10.4230/DagRep.10.2.50
Edited in cooperation with Franziska Eberle

1 Executive Summary

Nicole Megow (Universität Bremen, DE)
David Shmoys (Cornell University – Ithaca, US)
Ola Svensson (EPFL – Lausanne, CH)

This seminar was the sixth in a series of Dagstuhl "Scheduling" seminars (since 2008). Scheduling is a major research field that is studied from a practical and theoretical perspective in computer science, mathematical optimization, and operations research. Applications range from traditional production scheduling and project planning to the newly arising resource management tasks in the advent of internet technology and shared resources.

This edition of the seminar focused on the interplay between scheduling problems and problems that arise in the management of traffic. There are several notable aspects of the scheduling problems that arise particularly in this context:

- the role of dynamic decision-making in which data-driven approaches emerge (especially those that have stochastic elements in modelling multi-stage decision-making);
- the interplay between scheduling aspects and what might be viewed as routing aspects, providing a spacial component to the nature of the scheduling problem;
- the tension between questions of coordination and competition that arise from the fact that, for many of the issues in this domain, there are significant questions that depend on the extent to which the traffic can be centrally coordinated.

Scheduling, *Dagstuhl Reports*, Vol. 10, Issue 2, pp. 50–75
Editors: Nicole Megow, David Shmoys, and Ola Svensson
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Since the community working on the intersection of scheduling and traffic is itself rather broad, the seminar focused on researchers whose methodological focus relies on tools from the theoretical design of algorithms, on mathematical optimization methods, and on the combination of optimization and game-theoretic approaches.

Organization of the Seminar. The workshop brought together 59 researchers from theoretical computer science, mathematical optimization and operations research. The participants consisted of both senior and junior researchers, including a number of postdocs and advanced PhD students.

During the five days of the workshop, 31 talks of different lengths took place. Four keynote speakers gave an overview of the state-of-the art of the respective area in 60 minutes:

- Shuchi Chawla: Mechanisms for resource allocation
- Benjamin Moseley: Combinatorial Optimization Augmented with Machine Learning
- Evá Tardos: Learning in Games and in Queueing Systems
- Vera Traub: Approximation algorithms for traveling salesman problems.

The remaining slots were filled with shorter talks of 30 minutes on various topics related to scheduling, routing, transportation, mechanism design, learning, and applications in practice. Another highlight of the workshop was a historical note given by Jan Karel Lenstra with his view on the dynamic development of the area of scheduling in the past 60 years. Further, in the beginning of the week, open problem sessions were held. Throughout the week, a few sessions with spotlight talks of 8 minutes gave participants the chance to announce recent results and invite for discussions. The schedule left ample free time that was actively used for fruitful discussions and joint research.

Outcome. Organizers and participants regard the workshop as a great success. The workshop achieved the goal to bring together the related communities, share the state-of-the art research and discuss the current major challenges. The talks were excellent and very stimulating; participants actively met in working groups in the afternoon and evenings. It was remarked very positively that a significant number of younger researchers (postdocs and PhD students) participated and integrated very well.

The organizers wish to express their gratitude towards the Scientific Directorate and the administration of the Dagstuhl Center for their great support for this workshop.

2 Table of Contents

Executive Summary

Overview of Talks

Open Problems

3 Overview of Talks

3.1 Minimizing Energy Consumption on Multiple Machines with Sleep States

Antonios Antoniadis (MPI für Informatik – Saarbrücken, DE)

Joint work of Antonios Antoniadis, Naveen Garg, Gunjan Kumar, Nikhil Kumar
Main reference Antonios Antoniadis, Naveen Garg, Gunjan Kumar, Nikhil Kumar: "Parallel Machine Scheduling to Minimize Energy Consumption", in Proc. of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020, pp. 2758–2769, SIAM, 2020.
URL http://dx.doi.org/10.1137/1.9781611975994.168

The talk is about the problem of minimizing energy consumption on multiple machines with sleep states: We are given n jobs, each with a release date, a deadline and a processing time, and wish to schedule them on m parallel machines so as to minimize the total energy consumed. Machines can enter a sleep state and they consume no energy in this state. Each machine requires L units of energy to awaken from the sleep state and in its active state the machine can process jobs and consumes a unit of energy per unit time.

The core of the talk revolves around giving a 2-approximation algorithm for the single machine case. The algorithm is based on the solution of a linear programming relaxation, which can be decomposed into a convex combination of integer solutions. However none of them may be feasible since the linear programming relaxation has a strictly positive integrality gap. We discuss how such an integer solution can nevertheless be turned into a feasible solution without increasing the cost by too much. We conclude the talk by outlining how these ideas can be extended in order to obtain the first known constant approximation for the multiprocessor setting (with an approximation factor of 3).

3.2 General Framework for Metric Optimization Problems with Delay or with Deadlines

Yossi Azar (Tel Aviv University, IL)

Joint work of Yossi Azar, Noam Touitou
Main reference Yossi Azar, Noam Touitou: "General Framework for Metric Optimization Problems with Delay or with Deadlines", in Proc. of the 60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019, pp. 60–71, IEEE Computer Society, 2019.
URL https://doi.org/10.1109/FOCS.2019.00013

In this paper, we present a framework used to construct and analyze algorithms for online optimization problems with deadlines or with delay over a metric space. Using this framework, we present algorithms for several different problems. We present an $O(D^2)$ -competitive deterministic algorithm for online multilevel aggregation with delay on a tree of depth D, an exponential improvement over the $O(D^4 2^D)$ -competitive algorithm of Bienkowski et al. (ESA '16), where the only previously-known improvement was for the special case of deadlines by Buchbinder et al. (SODA '17). We also present an $O(\log^2 n)$ -competitive randomized algorithm for online service with delay over any general metric space of n points, improving upon the $O(\log^4 n)$ -competitive algorithm by Azar et al. (STOC '17). In addition, we present the problem of online facility location with deadlines. In this problem, requests arrive over time in a metric space, and need to be served until their deadlines by facilities that are opened momentarily for some cost. We also consider the problem of facility location with delay, in which the deadlines are replaced with arbitrary delay functions. For those problems, we present $O(\log^2 n)$ -competitive algorithms, with n the number of points in the metric space. The algorithmic framework we present includes techniques for the design of algorithms as well as techniques for their analysis.

3.3 Learning Augmented Energy Minimization via Speed Scaling

Etienne Bamas (EPFL – Lausanne, CH)

As power management has become a primary concern in modern data centers, computing resources are being scaled dynamically to minimize energy consumption. A classic speed scaling problem of Yao et al. (FOCS 1995) consists in scaling the speed of a processor dynamically to minimize energy under the constraint of finishing all the jobs in-between their release time and their deadline. This problem has been well studied both in offline and online settings. In this work, we initiate the study of the online speed scaling problem in a new framework in which machine learning predictions about the future can be integrated naturally. Inspired by recent work on learning-augmented online algorithms, we propose an algorithm which incorporates predictions in a black-box manner and outperforms any online algorithm if the accuracy is high, yet maintains provable guarantees if the prediction is very inaccurate.

3.4 Mechanisms for resource allocation

Shuchi Chawla (University of Wisconsin – Madison, US)

In this talk we will consider the problem of allocating multiple items to buyers who have values/preferences for subsets of items. We will consider a strategic setting where buyers can misreport their values so as to get as large an allocation as cheaply as possible. The algorithm knows the distributions from which buyers draw their values, but not the actual values themselves. Our goal is to design truthful mechanisms. I will talk about two objectives: social welfare maximization, where we want to maximize the sum of all buyers' values from their allocations; and revenue maximization, where we want to maximize the total payment the mechanism receives from all the buyers. In some contexts we will consider the online version of the problem.

This is a survey talk. I will describe what is known for the above problems under different assumptions on the instances, and what the main open problems are. The emphasis will be on algorithmic rather than economic issues.

3.5 The online k-taxi problem

Christian Coester (CWI – Amsterdam, NL)

License Creative Commons BY 3.0 Unported license
© Christian Coester
Joint work of Christian Coester, Elias Koutsoupias
Main reference Christian Coester, Elias Koutsoupias: "The online k-taxi problem", in Proc. of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019, pp. 1136–1147, ACM, 2019.
URL http://dx.doi.org/10.1145/3313276.3316370

In the k-taxi problem, a generalization of the k-server problem, an algorithm controls k taxis to serve a sequence of requests in a metric space. A request consists of two points s and t, representing a passenger that wants to be carried by a taxi from s to t. For each request and without knowledge of future requests, the algorithm has to select a taxi to transport the passenger. The goal is to minimize the total distance traveled by all taxis. The problem comes in two flavors, called the easy and the hard k-taxi problem: In the easy k-taxi problem, the cost is defined as the total distance traveled by the taxis; in the hard k-taxi problem, the cost is only the distance of empty runs.

The easy k-taxi problem is exactly equivalent to the k-server problem. The talk will focus mostly on the hard version, which is substantially more difficult. For hierarchically separated trees, I will present a memoryless randomized algorithm with optimal competitive ratio $2^k - 1$. This implies an $O(2^k \log n)$-competitive algorithm for arbitrary n-point metrics, which is the first competitive algorithm for the hard k-taxi problem for general finite metrics and general k. I will also describe main ideas of an algorithm based on growing, shrinking and shifting regions which achieves a constant competitive ratio for three taxis on the line (abstracting the scheduling of three elevators).

3.6 A Tale of Santa Claus, Hypergraphs and Matroids

Sami Davies (University of Washington – Seattle, US)

License Creative Commons BY 3.0 Unported license
© Sami Davies
Joint work of Sami Davies, Thomas Rothvoss, and Yihao Zhang
Main reference Sami Davies, Thomas Rothvoss, Yihao Zhang: "A Tale of Santa Claus, Hypergraphs and Matroids", in Proc. of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020, pp. 2748–2757, SIAM, 2020.
URL http://dx.doi.org/10.1137/1.9781611975994.167

A well-known problem in scheduling and approximation algorithms is the Santa Claus problem. Suppose that Santa Claus has a set of gifts, and he wants to distribute them among a set of children so that the least happy child is made as happy as possible. Here, the value that a child i has for a present j is of the form $p_{ij} \in \{0, p_j\}$. A polynomial time algorithm by Annamalai et al. gives a 12.33-approximation and is based on a modification of Haxell's hypergraph matching argument. In this paper, we introduce a matroid version of the Santa Claus problem. Our algorithm is also based on Haxell's augmenting tree, but with the introduction of the matroid structure we solve a more general problem with cleaner methods. Our result can then be used as a blackbox to obtain a $(4 + \varepsilon)$-approximation for Santa Claus. This factor also compares against a natural, compact LP for Santa Claus.

3.7 Optimally Handling Commitment Issues in Online Throughput Maximization

Franziska Eberle (Universität Bremen, DE)

Joint work of Franziska Eberle, Nicole Megow, and Kevin Schewior
Main reference Franziska Eberle, Nicole Megow, Kevin Schewior: "Optimally handling commitment issues in online throughput maximization", CoRR, Vol. abs/1912.10769, 2019.
URL https://arxiv.org/abs/1912.10769

We consider a fundamental online scheduling problem in which jobs with processing times and deadlines arrive online over time at their release dates. The task is to determine a feasible preemptive schedule on m machines that maximizes the number of jobs that complete before their deadline. Due to strong impossibility results for competitive analysis, it is commonly required that jobs contain some *slack* $\varepsilon > 0$, which means that the feasible time window for scheduling a job is at least $1 + \varepsilon$ times its processing time. In this paper, we answer the question on how to handle commitment requirements which enforce that a scheduler has to guarantee at a certain point in time the completion of admitted jobs. This is very relevant, e.g., in providing cloud-computing services and disallows last-minute rejections of critical tasks. We present the first online algorithm for handling commitment on parallel machines for small slack ε. When the scheduler must commit upon starting a job, the algorithm is $\Theta(\frac{1}{\varepsilon})$-competitive. Somewhat surprisingly, this is the same optimal performance bound (up to constants) as for scheduling without commitment on a single machine. If commitment decisions must be made before a job's slack becomes less than a δ-fraction of its size, we prove a competitive ratio of $O(\frac{\varepsilon}{\delta(\varepsilon--\delta)})$ for $0 < \delta < \varepsilon$. This result nicely interpolates between commitment upon starting a job and commitment upon arrival. For the latter commitment model, it is known that no (randomized) online algorithms does admit any bounded competitive ratio.

3.8 Pricing in Resource Allocation Games Based on Duality Gaps

Tobias Harks (Universität Augsburg, DE)

Main reference Tobias Harks: "Pricing in Resource Allocation Games Based on Duality Gaps", CoRR, Vol. abs/1907.01976, 2019.
URL http://arxiv.org/abs/1907.01976

We consider a basic resource allocation game, where the players' strategy spaces are subsets of R^m and cost/utility functions are parameterized by some common vector $u \in R^m$ and, otherwise, only depend on the own strategy choice. A strategy of a player can be interpreted as a vector of resource consumption and a joint strategy profile naturally leads to an aggregate consumption vector. Resources can be priced, that is, the game is augmented by a price vector $\lambda \in R^m_+$ and players have quasi-linear overall costs/utilities meaning that in addition to the original costs/utilities, a player needs to pay the corresponding price per consumed unit. We investigate the following question: for which aggregated consumption vectors u can we find prices λ that induce an equilibrium realizing the targeted consumption profile?

For answering this question, we revisit a well-known duality-based framework and derive several characterizations of the existence of such u and λ. We show that the characterization can help to unify parts of three largely independent streams in the literature – tolls in transportation systems, Walrasian market equilibria and congestion control in communication networks. Besides reproving existing results we establish novel existence results by using methods from polyhedral combinatorics, global optimization and discrete convexity.

3.9 Scheduling stochastic jobs with release dates on a single machine

Sven Jäger (TU Berlin, DE)

We consider the problem of minimizing the expected sum of weighted job completion times when jobs have stochastic processing times and may arrive over time. While in the offline model, all jobs are known upfront, jobs are only revealed at their release dates in the online model. The problem on identical parallel machines has been considered by Möhring, Schulz, and Uetz (1999), Megow, Uetz, and Vredeveld (2006), and Schulz (2008). Möhring, Schulz, and Uetz observed that their offline policy has a performance guarantee of 3 in the case of a single machine. A refined analysis of the policies developed by Schulz shows that in the single-machine case they are a 2.619-competitive deterministic online policy and a 2-competitive randomized online policy. These are also the best known performance guarantees of any offline policy. In the talk I will sketch how to obtain a (randomized) competitive ratio below 2 in the special case of NBUE processing times. This is based on a similar analysis as by Goemans, Queyranne, Schulz, Skutella, and Wang (2002) for the problem with deterministic processing times.

3.10 Online Learning with Vector Costs and Bandits with Knapsacks

Thomas Kesselheim (Universität Bonn, DE)

Joint work of Thomas Kesselheim , Sahil Singla

We introduce an online learning problem with vector costs (OLVC). Akin to online generalized load balancing, in each time step an algorithm chooses one of n actions and then incurs a vector cost $[0, 1]^d$, which depends on the chosen action. The goal of the online algorithm is to minimize the ℓ_p norm of the sum of its cost vectors. The difference is that incurred costs are not known until after having chosen the action. This way, the setting generalizes the classical online learning setting, which is captured by $d = 1$.

We study OLVC in both stochastic and adversarial arrival settings, and give a general procedure to reduce the problem from d dimensions to a single dimension. This allows us to use classical online learning algorithms in both full and bandit feedback models to obtain (near) optimal results. In particular, we obtain a single algorithm (up to the choice of learning rate) that gives sublinear regret for stochastic arrivals and a tight $O(\min\{p, \log d\})$ competitive ratio for adversarial arrivals.

The OLVC problem also occurs as a natural subproblem when trying to solve the popular Bandits with Knapsacks (BWK) problem. This connection allows us to use our OLVCp techniques to obtain (near) optimal results for BWK in both stochastic and adversarial settings. In particular, we obtain a tight $O(\log d \cdot \log T)$ competitive ratio algorithm for adversarial BWK, which improves over the $O(d \cdot \log T)$ competitive ratio algorithm of Immorlica et al. (2019).

3.11 Equilibria in Atomic Splittable Congestion Games

Max Klimm (HU Berlin, DE)

Joint work of Max Klimm, Philipp Warode
Main reference Max Klimm, Philipp Warode: “Complexity and Parametric Computation of Equilibria in Atomic Splittable Congestion Games via Weighted Block Laplacians”, in Proc. of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020, pp. 2728–2747, SIAM, 2020.
URL httpd://doi.org/10.1137/1.9781611975994.166

We settle the complexity of computing an equilibrium in atomic splittable congestion games with player-specific affine cost functions showing that it is PPAD-complete. To prove that the problem is contained in PPAD, we develop a homotopy method that traces an equilibrium for varying flow demands of the players. A key technique is to describe the evolution of the equilibrium locally by a novel block Laplacian matrix. This leads to a path following formulation where states correspond to supports that are feasible for some demands and neighboring supports are feasible for increased or decreased flow demands. A closer investigation of the block Laplacian system allows to orient the states giving rise to unique predecessor and successor states thus putting the problem into PPAD. For the PPAD-hardness, we reduce from computing an approximate equilibrium of a bimatrix win-lose game. As a byproduct of our reduction we further show that computing a multiclass Wardrop equilibrium with class-dependent affine cost functions is PPAD-complete as well. As a byproduct of our PPAD-completeness proof, we obtain an algorithm that computes all equilibria parametrized by the players' flow demands. For player-specific costs, this computation may require several increases and decreases of the demands leading to an algorithm that runs in polynomial space but exponential time. For player-independent costs only demand increases are necessary. If the coefficients be,i are in general position, this yields an algorithm computing all equilibria as a function of the flow demand running in time polynomial in the output.

3.12 Non-Clairvoyant Precedence Constrained Scheduling

Amit Kumar (Indian Institute of Technology – New Dehli, IN)

Joint work of Naveen Garg, Anupam Gupta, Amit Kumar, Sahil Singla
Main reference Naveen Garg, Anupam Gupta, Amit Kumar, Sahil Singla: “Non-Clairvoyant Precedence Constrained Scheduling”, in Proc. of the 46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece, LIPIcs, Vol. 132, pp. 63:1–63:14, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019.
URL https://doi.org/10.4230/LIPIcs.ICALP.2019.63

We consider the online problem of scheduling jobs on identical machines, where jobs have precedence constraints. We are interested in the demanding setting where the jobs sizes are not known up-front, but are revealed only upon completion (the non-clairvoyant setting). Such precedence-constrained scheduling problems routinely arise in map-reduce and large-scale optimization. For minimizing the total weighted completion time, we give a constant-competitive algorithm. And for total weighted flow-time, we give an $O(1/\epsilon^2)$-competitive algorithm under $(1 + epsilon)$-speed augmentation and a natural “"no-surprises” assumption on release dates of jobs (which we show is necessary in this context). Our algorithm proceeds by assigning virtual rates to all waiting jobs, including the ones which are dependent on other uncompleted jobs. We then use these virtual rates to decide on the actual rates of minimal jobs (i.e., jobs which do not have dependencies and hence are eligible to run). Interestingly, the virtual rates are obtained by allocating time in a fair manner, using a Eisenberg-Gale-type convex program (which we can solve optimally using a primal-dual scheme). The optimality condition of this convex program allows us to show dual-fitting proofs more easily, without having to guess and hand-craft the duals. This idea of using fair virtual rates may have broader applicability in scheduling problems.

3.13 Online Vehicle Routing, the edge of optimization in large scale applications

Sebastien Martin (LYFT – New York, US)

Joint work of Dimitris Bertsimas, Patrick Jaillet, and Sébastien Martin
Main reference Dimitris Bertsimas, Patrick Jaillet, Sébastien Martin: "Online Vehicle Routing: The Edge of Optimization in Large-Scale Applications", Oper. Res., Vol. 67(1), pp. 143–162, 2019.
URL https://doi.org/10.1287/opre.2018.1763

With the emergence of ride-sharing companies that offer transportation on demand at a large scale and the increasing availability of corresponding demand data sets, new challenges arise to develop routing optimization algorithms that can solve massive problems in real time. In this paper, we develop an optimization framework, coupled with a novel and generalizable backbone algorithm, that allows us to dispatch in real time thousands of taxis serving more than 25,000 customers per hour. We provide evidence from historical simulations using New York City routing network and yellow cab data to show that our algorithms improve upon the performance of existing heuristics in such real-world settings.

3.14 Malleable Scheduling Beyond Identical Machines

Jannik Matuschke (KU Leuven, BE)

Joint work of Dimitris Fotakis, Jannik Matuschke, Orestis Papadigenopoulos
Main reference Dimitris Fotakis, Jannik Matuschke, Orestis Papadigenopoulos: "Malleable Scheduling Beyond Identical Machines", in Proc. of the Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2019, September 20-22, 2019, Massachusetts Institute of Technology, Cambridge, MA, USA, LIPIcs, Vol. 145, pp. 17:1–17:14, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019.
URL http://dx.doi.org/10.4230/LIPIcs.APPROX-RANDOM.2019.17

In malleable scheduling, jobs can be executed simultaneously on multiple machines with the processing time depending on the number of allocated machines. Each job is required to be executed non-preemptively and in unison, i.e., it has to occupy the same time interval on all its allocated machines. In this talk, we discuss a generalization of malleable scheduling, in which a function $f(S, j)$ determines the processing time of job j on machine subset S. We derive a constant factor approximation for minimizing the makespan in the case that $f(S, j)$ can be expressed in terms of job-dependent machine speeds and fulfills a non-decreasing workload assumption for each fixed job j. We also discuss further generalizations and open problems.

3.15 Combinatorial Optimization Augmented with Machine Learning

Benjamin J. Moseley (Carnegie Mellon University – Pittsburgh, US)

Combinatorial optimization often focuses on optimizing for the worst-case. However, the best algorithm to use depends on the "relative inputs", which is application specific and often does not have a formal definition.

The talk gives a new theoretical model for designing algorithms that are tailored to inputs for the application at hand. In the model, learning is performed on past problem instances to make predictions on future instances. These predictions are incorporated into the design and analysis of the algorithm. The predictions can be used to achieve "instance-optimal" algorithm design when the predictions are accurate and the algorithm's performance gracefully degrades when there is error in the prediction.

The talk will apply this framework to applications in online algorithm design and give algorithms with theoretical performance that goes beyond worst-case analysis. The majority of the talk will focus on load balancing on unrelated machines.

3.16 Group Fairness in Network Design and Combinatorial Optimization

Kamesh Munagala (Duke University – Durham, US)

Joint work of Yu Cheng, Zhihao Jiang, Kamesh Munagala, Kangning Wang
Main reference Yu Cheng, Zhihao Jiang, Kamesh Munagala, Kangning Wang: "Group Fairness in Committee Selection", in Proc. of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019, pp. 263–279, ACM, 2019.
URL https://doi.org/10.1145/3328526.3329577

Consider the following classical network design model. There are n clients in a multi-graph with a single sink node. Each edge has a cost to buy, and a length if bought; typically, costlier edges have smaller lengths. There is a budget B on the total cost of edges bought. Given a set of bought edges, the distance of a client to the sink is the shortest path according to the edge lengths. Such a model captures buy-at-bulk network design and facility location as special cases.

Rather than pose this as a standard optimization problem, we ask a different question: Suppose a provider is allocating budget B to build this network, how should it do so in a manner that is fair to the clients? We consider a classical model of group fairness termed the core in cooperative game theory: If each client contributes its share B/n amount of budget as tax money, no subset of clients should be able to pool their tax money to build a different network that simultaneously improves all their distances to the sink. The question is: Does such a solution always exist, or approximately exist?

We consider an abstract "committee selection" model from social choice literature that captures not only the above problem, but other combinatorial optimization problems where we need to provision public resources subject to combinatorial constraints, in order to provide utility to clients. For this general model, we show that an approximately fair solution always exists, where the approximation scales down the tax money each client can use for deviation by only a constant factor. Our existence result relies on rounding an interesting fractional relaxation to this problem. In certain cases such as the facility location problem, it also implies a polynomial time algorithm. We conclude with several open questions.

3.17 Scheduling Bidirectional Traffic

Rolf H. Möhring (TU Berlin, DE)

Joint work of Elisabeth Lübbecke, Marco Lübbecke, Rolf H. Möhring
Main reference Elisabeth Lübbecke, Marco E. Lübbecke, Rolf H. Möhring: "Ship Traffic Optimization for the Kiel Canal", Oper. Res., Vol. 67(3), pp. 791–812, 2019.
URL https://doi.org/10.1287/opre.2018.1814

We introduce, discuss, and solve a hard practical optimization problem that deals with routing bidirectional traffic. This situation occurs in train traffic on a single track with sidings, ship traffic in a canal, or bidirectional data communication. We have developed a combinatorial algorithm that provides a unified view of routing and scheduling that combines joint (global) and sequential (local) solution approaches to allocate scarce network resources to a stream of online arriving vehicles in a collision-free manner. Computational experiments on real traffic data with results obtained by human expert planners show that our algorithm improves upon manual planning by 25%.

This combination of routing and scheduling leads to a new class of scheduling problems, and we will also address some complexity and approximation results for this class.

3.18 Stochastic Makespan Minimization

Viswanath Nagarajan (University of Michigan – Ann Arbor, US)

Joint work of Anupam Gupta, Amit Kumar, Viswanath Nagarajan, Xiangkun Shen
Main reference Anupam Gupta, Amit Kumar, Viswanath Nagarajan, Xiangkun Shen: "Stochastic Makespan Minimization in Structured Set Systems (Extended Abstract)", in Proc. of the Integer Programming and Combinatorial Optimization – 21st International Conference, IPCO 2020, London, UK, June 8-10, 2020, Proceedings, Lecture Notes in Computer Science, Vol. 12125, pp. 158–170, Springer, 2020.
URL http://dx.doi.org/10.1007/978-3-030-45771-6_13

We consider stochastic combinatorial optimization problems where the objective is to minimize the expected makespan. First, we provide a constant-factor approximation algorithm for stochastic makespan minimization on unrelated machines. Second, we provide an $O(\log \log m)$ approximation algorithm for stochastic resource allocation problems with some geometric structure, such as intervals in a line, paths in a tree and rectangles/disks in the plane. Both results utilize (i) an exponential-size LP based on the cumulant generating function and (ii) an iterative rounding algorithm.

3.19 Online Algorithms via Projections

Seffi Naor (Technion – Haifa, IL)

Joint work of Niv Buchbinder, Anupam Gupta, Marco Molinaro, Joseph (Seffi) Naor
Main reference Niv Buchbinder, Anupam Gupta, Marco Molinaro, Joseph (Seffi) Naor: "k-Servers with a Smile: Online Algorithms via Projections", in Proc. of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019, pp. 98–116, SIAM, 2019.
URL https://doi.org/10.1137/1.9781611975482.7

We present a new/old approach to the design of online algorithms via Bregman projections. This approach is applicable to a wide range of online problems and we discuss connections to previous work on online primal-dual algorithms. In particular, the k-server problem on trees and HSTs is considered. The projection-based algorithm for this problem turns out to have a competitive ratio that matches some of the recent results given by Bubeck et al. (STOC 2018), whose algorithm uses mirror-descent-based continuous dynamics prescribed via a differential inclusion.

3.20 Improved Approximation Algorithms for Inventory Problems

Neil Olver (London School of Economics and Political Science, GB)

Joint work of Thomas Bosman, Neil Olver
Main reference Thomas Bosman, Neil Olver: "Improved Approximation Algorithms for Inventory Problems", in Proc. of the Integer Programming and Combinatorial Optimization – 21st International Conference, IPCO 2020, London, UK, June 8-10, 2020, Proceedings, Lecture Notes in Computer Science, Vol. 12125, pp. 91–103, Springer, 2020.
URL https://doi.org/10.1007/978-3-030-45771-6_8

We give new approximation algorithms for the submodular joint replenishment problem and the inventory routing problem, using an iterative rounding approach. In both problems, we are given a set of N items and a discrete time horizon of T days in which given demands for the items must be satisfied. Ordering a set of items incurs a cost according to a set function, with properties depending on the problem under consideration. Demand for an item at time t can be satisfied by an order on any day prior to t, but a holding cost is charged for storing the items during the intermediate period; the goal is to minimize the sum of the ordering and holding cost. Our approximation factor for both problems is $O(\log \log \min(N, T))$; this improves exponentially on the previous best results.

3.21 Sample-Based Prophet Inequalities

Kevin Schewior (Universität Köln, DE)

Joint work of José R. Correa, Paul Dütting, Felix A. Fischer, and Kevin Schewior
Main reference José R. Correa, Paul Dütting, Felix A. Fischer, Kevin Schewior: "Prophet Inequalities for I.I.D. Random Variables from an Unknown Distribution", in Proc. of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019, pp. 3–17, ACM, 2019.
URL https://doi.org/10.1145/3328526.3329627

Consider a gambler who observes independent draws from a known sequence of positive-valued distributions. Upon observation of any value, the gambler has to decide whether to keep it as final reward or to discard it forever. The prophet inequality (Krengel, Sucheston and Garling 1978) states that there is a strategy whose expected accepted value is at least half of the expected maximum of all draws (the prophet's value). We first review the recent result (Wang 2018) that the following simple strategy also achieves the same ratio: Sample one value from each of the distributions, and set their maximum as a threshold for accepting any value.

We then turn to the case in which all distributions are identical. It is a simple corollary from results on the secretary problem that a ratio of $1/e$ is achievable without samples (again with respect to the expected maximum). We show using Ramsey's Theorem that this is best possible. We also show that knowing $O(n^2)$ samples is essentially as good as knowing the distribution, meaning that a ratio of 0.745 can be approached in that case (Correa et al. 2017). For the remainder of the talk, we work towards understanding the case with $O(n)$ samples, but, unlike for the distinct-distributions case, open questions remain.

3.22 Online Metric Algorithms with Untrusted Predictions

Bertrand Simon (Universität Bremen, DE)

Joint work of Antonios Antoniadis, Christian Coester, Marek Eliás, Adam Polak, and Bertrand Simon
Main reference Antonios Antoniadis, Christian Coester, Marek Eliás, Adam Polak, Bertrand Simon: "Online metric algorithms with untrusted predictions", CoRR, Vol. abs/2003.02144, 2020.
URL https://arxiv.org/abs/2003.02144

Machine-learned predictors, although achieving very good results for inputs resembling training data, cannot possibly provide perfect predictions in all situations. Still, decision-making systems that are based on such predictors need not only to benefit from good predictions but also to achieve a decent performance when the predictions are inadequate. In this paper, we propose a prediction setup for Metrical Task Systems (MTS), a broad class of online decision-making problems including, e.g., caching, k-server and convex body chasing. We utilize results from the theory of online algorithms to show how to make the setup robust. We extend our setup in two ways, (1) adapting it beyond MTS to the online matching on the line problem, and (2) specifically for caching, to achieve an improved dependence on the prediction error. Finally, we present an empirical evaluation of our methods on real world datasets, which suggests practicality.

3.23 Fixed-Order Scheduling on Parallel Machines

René Sitters (VU University of Amsterdam, NL)

Joint work of René Sitters, Thomas Bosman, Dario Frascaria, Neil Olver, René Sitters, and Leen Stougie
Main reference Thomas Bosman, Dario Frascaria, Neil Olver, René Sitters, Leen Stougie: "Fixed-Order Scheduling on Parallel Machines", in Proc. of the Integer Programming and Combinatorial Optimization – 20th International Conference, IPCO 2019, Ann Arbor, MI, USA, May 22-24, 2019, Proceedings, Lecture Notes in Computer Science, Vol. 11480, pp. 88–100, Springer, 2019.
URL https://doi.org/10.1007/978-3-030-17953-3_7

We consider the following natural scheduling problem: Given a sequence of jobs with weights and processing times, one needs to assign each job to one of m identical machines in order to minimize the sum of weighted completion times. The twist is that for machine the jobs assigned to it must obey the order of the input sequence, as is the case in multi-server queuing systems. We establish a constant factor approximation algorithm for this (strongly NP-hard) problem.

3.24 Approximation Algorithms for Replenishment Problems with Fixed Turnover Times

Leen Stougie (CWI – Amsterdam, NL)

Joint work of Thomas Bosman, Martijn van EE, Yang Jiao, Alberto Marchetti Spaccamela, R Ravi, Leen Stougie
Main reference Thomas Bosman, Martijn van Ee, Yang Jiao, Alberto Marchetti-Spaccamela, R. Ravi, Leen Stougie: "Approximation Algorithms for Replenishment Problems with Fixed Turnover Times", in Proc. of the LATIN 2018: Theoretical Informatics – 13th Latin American Symposium, Buenos Aires, Argentina, April 16-19, 2018, Proceedings, Lecture Notes in Computer Science, Vol. 10807, pp. 217–230, Springer, 2018.
URL https://doi.org/10.1007/978-3-319-77404-6_17

We introduce and study a class of optimization problems we call replenishment problems with fixed turnover times: a very natural model that has received little attention in the literature. Clients with capacity for storing a certain commodity are located at various places; at each client the commodity depletes within a certain time, the turnover time, which is constant but can vary between locations. Clients should never run empty. The natural feature that makes this problem interesting is that we may schedule a replenishment (well) before a client becomes empty, but then the next replenishment will be due earlier also. This added workload needs to be balanced against the cost of routing vehicles to do the replenishments. In this paper, we focus on the aspect of minimizing routing costs. However, the framework of recurring tasks, in which the next job of a task must be done within a fixed amount of time after the previous one is much more general and gives an adequate model for many practical situations. Note that our problem has an infinite time horizon. However, it can be fully characterized by a compact input, containing only the location of each client and a turnover time. This makes determining its computational complexity highly challenging and indeed it remains essentially unresolved. We study the problem for two objectives: min-avg minimizes the average tour cost and min-max minimizes the maximum tour cost over all days. For min-max we derive a logarithmic factor approximation for the problem on general metrics and a 6-approximation for the problem on trees, for which we have a proof of NP-hardness. For min-avg we present a logarithmic factor approximation on general metrics, a 2-approximation for trees, and a pseudopolynomial time algorithm for the line. Many intriguing problems remain open.

In this lecture I will explain the model and the complexity issues and give an intuitive idea of the approximation results on a tree.

3.25 Learning in Games and in Queueing Systems

Éva Tardos (Cornell University – Ithaca, US)

Over the last two decades we have developed good understanding how to quantify the impact of strategic user behavior on overall performance in many games (including traffic routing as well as online auctions), and showed that the resulting bounds extend to repeated games assuming players use a form of no-regret learning that helps them adapt to the environment. In this talk we will review these results, and study this phenomenon in the context of a game modeling queuing systems: routers compete for servers, where packets that do not get service will be resent at future rounds, resulting in a system where the number of packets at each round depends on the success of the routers in the previous rounds. In joint work with Jason Gaitonde, we analyze the resulting highly dependent random process and find that if the capacity of the servers is high enough to allow a centralized and knowledgeable scheduler to get all packets served even with double the packet arrival rate, then learning can help the queues in coordinating their behavior, the expected number of packets in the queues will remain bounded throughout time, assuming older packets have priority.

3.26 Approximation algorithms for traveling salesman problems

Vera Traub (Universität Bonn, DE)

In the traveling salesman problem we are given a finite set of cities with pairwise non-negative distances. The task is to find a shortest tour that visits all cities and returns to the starting point. The distances between cities can either be symmetric (TSP) or asymmetric (ATSP). We will also consider the path version, which is the generalization of the traveling salesman problem in which the endpoints of the tour are given and distinct.

For most of these traveling salesman problems improved approximation algorithms have been found during the past few years. The only exception is the symmetric TSP, where Christofides' classical 3/2-approximation from the 70's remains the best known approximation algorithm.

In this talk we survey the recent progress on approximation algorithms for both the symmetric and the asymmetric traveling salesman problem, as well as their path versions.

3.27 Greed…Is Good For Scheduling Under Uncertainty (2nd ed.)

Marc Uetz (University of Twente – Enschede, NL)

Joint work of Varun Gupta, Benjamin Moseley, Marc Uetz, Qiaomin Xie
Main reference Varun Gupta, Benjamin Moseley, Marc Uetz, Qiaomin Xie: "Greed Works—Online Algorithms for Unrelated Machine Stochastic Scheduling", Mathematics of Operations Research, Vol. 45(2), pp. 497–516, 2020.
URL https://doi.org/10.1287/moor.2019.0999

This spotlight talk reports about an update of a 2017 IPCO paper together with Varun Gupta, Ben Moseley, and Qiaomin Xie. In that updated paper, we show that a rather simple and intuitive greedy algorithm performs surprisingly well for a classical scheduling problem, namely minimizing the total weighted completion time on unrelated machines. In fact we give the first results for this problem when jobs are allowed to appear online over time, and have uncertain job sizes. Due to recent simplifications and improvements in both algorithm and analysis, our new performance bounds even improve upon previously best known results for the deterministic online problem, from the 1990s. The algorithm's basic idea is a greedy assignment of jobs to machines, just mimicking a "nominal" schedule where stochastic processing times are replaced by expectations. Moreover, the main idea for making this algorithm also competitive for uncertain job sizes, is is to adhere as much as possible to that nominal schedule. The analysis is based on dual fitting.

3.28 Nash flows over Time with Spillback and Kinematic Waves

Laura Vargas Koch (RWTH Aachen, DE)

Joint work of Leon Sering and Laura Vargas Koch
Main reference Leon Sering, Laura Vargas Koch: "Nash Flows Over Time with Spillback", in Proc. of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019, pp. 935–945, SIAM, 2019.
URL https://doi.org/10.1137/1.9781611975482.57

Modeling traffic in road networks is a widely studied but challenging problem, especially under the assumption that drivers act selfishly. A common approach is the deterministic queuing model, for which the structure of dynamic equilibria has been studied extensively in the last couple of years. The basic idea is to model traffic by a continuous flow that travels over time through a network, in which the arcs are endowed with transit times and capacities. Whenever the flow rate exceeds the capacity the flow particles build up a queue. So far it was not possible to represent the real-world phenomena spillback and kinematic waves in this model. By introducing a storage capacity arcs can become full, and thus, might block preceding arcs, i.e., spillback occurs. Furthermore, we model kinematic waves by upstream moving flows over time representing the gaps between vehicles. We carry over the main results of the original model to our generalization, i.e., we characterize Nash flows over time by sequences of particular static flows, so-called spillback thin flows. Furthermore, we give a constructive proof for the existence.

3.29 A Water-Filling Primal-Dual Algorithm for Approximating Non-Linear Covering Problems

Jose Verschae (O'Higgins University – Rancagua, CL)

Joint work of Andrés Fielbaum, Ignacio Morales, José Verschae
Main reference Andrés Fielbaum, Ignacio Morales, José Verschae: "A Water-Filling Primal-Dual Algorithm for Approximating Non-Linear Covering Problems", CoRR, Vol. abs/1912.12151, 2019.
URL https://arxiv.org/abs/1912.12151

Obtaining strong linear relaxations of capacitated covering problems constitute a major technical challenge even for simple settings. For one of the most basic cases, the Knapsack-Cover (Min-Knapsack) problem, the relaxation based on *knapsack-cover inequalities* achieves an integrality gap of 2. These inequalities are exploited in more general problems, many of which admit primal-dual approximation algorithms.

Inspired by problems from power and transport systems, we introduce a general setting in which items can be taken fractionally to cover a given demand. The cost incurred by an item is given by an arbitrary non-decreasing function of the chosen fraction. We generalize the knapsack-cover inequalities to this setting an use them to obtain a $(2+\varepsilon)$-approximate primal-dual algorithm. Our procedure has a natural interpretation as a bucket-filling algorithm, which effectively balances the difficulties given by having different slopes in the cost functions: when some superior portion of an item presents a low slope, it helps to increase the priority with which the inferior portions may be taken. We also present a rounding algorithm with an approximation guarantee of 2.

We generalize our algorithm to the Unsplittable Flow-Cover problem on a line, also for the setting where items can be taken fractionally. For this problem we obtain a $(4+\varepsilon)$-approximation algorithm in polynomial time, almost matching the 4-approximation known for the classical setting.

3.30 Dynamic Approximate Maximum Independent Set of Intervals, Hypercubes and Hyperrectangles

Andreas Wiese (University of Chile, CL)

Joint work of Monika Henzinger, Stefan Neumann, Andreas Wiese
Main reference Monika Henzinger, Stefan Neumann, Andreas Wiese: "Dynamic Approximate Maximum Independent Set of Intervals, Hypercubes and Hyperrectangles", CoRR, Vol. abs/2003.02605, 2020.
URL https://arxiv.org/abs/2003.02605

Independent set is a fundamental problem in combinatorial optimization. While in general graphs the problem is essentially inapproximable, for many important graph classes there are approximation algorithms known in the offline setting. These graph classes include interval graphs and geometric intersection graphs, where vertices correspond to intervals/geometric objects and an edge indicates that the two corresponding objects intersect. We present the first dynamic approximation algorithms for independent set of intervals and geometric objects. They work in the fully dynamic model where in each update an interval/geometric object is inserted or deleted. Our algorithms are deterministic and have worst-case update times that are polylogarithmic for constant d and ϵ. We achieve the following approximation ratios:

- For independent set of intervals, we maintain $(1+\epsilon)$-approximate solutions for the unweighted and the weighted case.
- For independent set of d-dimensional hypercubes we maintain $(1+\epsilon)2^d$-approximate solutions in the unweighted case and $O(2^d)$-approximate solutions in the weighted case. Also, we show that for maintaining unweighted $(1+\epsilon)$-approximate solutions one needs polynomial update time for $d \geq 2$ if the ETH holds.
- For weighted d-dimensional hyperrectangles we present a dynamic algorithm with approximation ratio $(1+\epsilon)\log^{d-1} N$, assuming that the coordinates of all input hyperrectangles are in $[0,N]^d$ and each of their edges has length at least 1.

4 Open Problems

4.1 Deterministic min-cost matching with delays

Yossi Azar (Tel Aviv University, IL)

We are given a metric space. Requests arrive over time. Requests need to be matched in pairs. Pending requests can be matched at any time after their arrival. The goal is to minimize the cost of the matching (distance between matched requests) plus the total delay of all requests (the delay of a request is the time between its arrival until it is matched). Denote the size of the metric space by n and the number of requests by m.

Known Results

- $O(\log n)$ randomized competitive algorithm for arbitrary metric space [1, 2]
- A lower bound of $\Omega(\log n/\log\log n)$ for randomized algorithm (even for the metric space of n integer point $[0,n]$ on the line) [3]
- $O(m^{0.59})$ deterministic competitive algorithm and $O(n)$ deterministic competitive algorithms. No lower bounds better then the randomized lower bound are known. [4, 5, 6]

Open: Close the gap between $\log n$ and n (or $m^{0.59}$) for deterministic algorithms.

Variant (bi-chromatic): Requests are red or blue (suppliers vs costumers) and the matching is always between requests of different colors.

Known results: Similar results to the previous model except of a weaker lower bound of $\Omega(\sqrt{\log n/\log\log n})$ for the (randomized) competitive ratio [3,6].

Open: close the gap between $\log n$ and n (or $m^{0.59}$) for deterministic algorithms for the bi-chromatic case.

References

1 Yuval Emek, Shay Kutten, and Roger Wattenhofer *Online matching: haste makes waste!*. STOC 2016: 333-344.

2 Yossi Azar, Ashish Chiplunkar, and Haim Kaplan. *Polylogarithmic bounds on the competitiveness of min-cost perfect matching with delays*. SODA 2017: 1038-1050.

3 Itai Ashlagi, Yossi Azar, Moses Charikar, Ashish Chiplunkar, Ofir Geri, Haim Kaplan, Rahul M. Makhijani, Yuyi Wang, and Roger Wattenhofer. *Min-cost bipartite perfect matching with delays*. APPROX-RANDOM 2017: 1:1-1:20

4 Marcin Bienkowski, Artur Kraska, and Pawel Schmidt. *A match in time saves nine: Deterministic online matching with delays.* WAOA 2017: 132-146.

5 Marcin Bienkowski, Artur Kraska, Hsiang-Hsuan Liu, and Pawel Schmidt. *A primal-dual online deterministic algorithm for matching with delays.* WAOA 2018: 51-68.

6 Yossi Azar, Amit Jacob Fanani: *Deterministic Min-Cost Matching with Delays.* Theory Comput. Syst. 64(4): 572-592 (2020) and WAOA 2018: 21-35

4.2 On-line Routing

Sanjoy Baruah (Washington Univerity in St. Louis, US)

Consider the following on-line routing problem on graphs. Each edge e of a directed graph is characterized by two upper bounds on the delay that will be encountered upon traversing it: an upper bound $c_W(e)$ on the maximum delay under all circumstances, and a (smaller) upper bound $c_T(e)$ on the maximum delay one would encounter under all "typical" (i.e., non-pathological) circumstances. The actual delay that will be encountered upon traversing an edge is unknown prior to actually traversing that edge.

Given a source vertex s, a destination vertex t,and a delay bound D, the objective is to travel from s to t such that one is guaranteed to arrive at t within D time units of leaving s, whilst simultaneously minimizing the duration taken in doing so under all typical (non-pathological) circumstances.

Algorithms have previously been proposed [1] for obtaining optimal such routes; however it is unknown whether or not the actual number of edges in all such optimal routes is polynomial in the problem specification.

References

1 Agrawal, K and Baruah, S.: *Adaptive Real-Time Routing in Polynomial Time*, Proceedings of the IEEE Real-Time Systems Symposium (RTSS 2019), Hong Kong, December 2019. IEEE Computer Society Press.

4.3 Integrality Gap of the natural LP for $P2|\text{prec}|C_{\max}$

Xinrui Jia (EPFL – Lausanne, CH)

This is an open problem stated in Kulkarni et al. 2020.

The problem is whether the natural LP relaxation for $P2|\text{prec}|C_{\max}$ has a large integrality gap when raised to $o(\log n)$ levels of the Sherali-Adams hierarchy. In the paper, a different scheduling problem, denoted $1|r_j, d_j|\sum_j p_j U'_j$, was demonstrated to have a large integrality gap when raised to $o(\log n)$ levels. This is the scheduling problem on one machine where jobs have release times r_j, deadlines d_j, processing times p_j, and the objective is to schedule the jobs non-preemptively for p'_j units of time $0 \leq p'_j \leq p_j$, to minimize $\sum_{j\in J}(p_j - -p'_j)$.

The authors present a way to transform instances of $1|r_j, d_j| \sum_j p_j U_j'$ into instances of $P2|\text{prec}|C_{\max}$, and show that an $o(\log n)$-level Sherali-Adams lift does not lead to a $(1+\epsilon)$ approximation for $1|r_j, d_j| \sum_j p_j U_j'$. The authors believe that the same instance used is the right one for showing an integrality gap for $P2|\text{prec}|C_{\max}$.

References

1 Kulkarni J., Li S., Tarnawski J., and Ye M. *Hierarchy-Based Algorithms for Minimizing Makespan under Precedence and Communication Constraints.* SODA, 2020.

4.4 The Busy Time Problem

Samir Khuller (Northwestern University – Evanston, US)

Given a set of jobs with release times, deadlines and processing times. The goal is to partition the jobs into bundles such that each bundle consists of a set of jobs that can be executed non-preemptively on a batch machine with batch capacity B. The total cost of this schedule is the sum of lengths of all the bundles. Each job has to be scheduled respecting its release time and deadline. The goal is to find a minimum cost schedule. Since the number of bundles is unbounded, a feasible solution can simply put each job alone in a bundle, or any B jobs in a bundle.

We are interested in approximation algorithms for this problem. The paper below presents a fairly simple "Greedy tracking" algorithm for this problem with a factor 3 guarantee. The worst example known for this algorithm is 2, so its true performance might be better.

References

1 Chang J., Khuller S., and Mukherjee K. *LP rounding and combinatorial algorithms for minimizing active and busy time.*. J. Sched. 20(6): 657-680 (2017)

4.5 Location Routing with Depot Capacities

Jannik Matuschke (KU Leuven, BE)

In the classic Facility Location Problem, we are given a set of possible locations for facilities with associated opening costs and a set of clients. We have to decide on which facilities to open. Then each client is served by its closest open facility and we have to pay the sum of all opening and connection costs. A drawback of this classic problem is that it assumes every client to receive its own dedicated connection, whereas in practice, several nearby clients can easily be served by the same vehicle. Ignoring these synergies and thus over-estimating connection costs can lead to inferior solutions whose cost significantly exceeds that of an optimal solution. This motivates the study of Location Routing, a combination of Facility Location and Vehicle Routing, in which clients are served from tours originating at open facilities.

Capacitated Location Routing Problem

Input: a set of clients C, a set of facilities F, a metric d on $C \cup F$, opening costs costs f_i for each $i \in F$, a facility capacity $B \in \mathbb{Z}_+$, a vehicle capacity $U \in \mathbb{Z}_+$

Task: Find a set facility S and a set of tours[1] $\mathcal{T}$ such that

1. $C \subseteq \bigcup_{T \in \mathcal{T}} V(T)$ (every client is served by a tour),
2. $|V(T) \cap S| = 1$ for all $T \in \mathcal{T}$ (every tour contains an open facility),
3. $|V(T) \cap C| \leq U$ for all $T \in \mathcal{T}$ (every tour serves at most U clients),
4. $\bigcup_{T \in \mathcal{T} : i \in V(T)} |V(T) \cap C| \leq B$ for all $i \in S$ (every facility serves at most B clients),

minimizing the total cost $\sum_{i \in S} f_i + \sum_{T \in \mathcal{T}} d(T)$.

Open question: Is there a constant factor approximation algorithm for the Capacitated Location Routing Problem?

Known results. For the case that facilities are uncapacitated ($B = \infty$), a constant factor approximation is known [1, 2], based on the combination of two lower bounds derived from instances of Facility Location and Minimum Spanning Tree, respectively. For the general problem with capacitated facilities, similar lower bounds can be combined with the classic LP rounding approach for unrelated machine scheduling [3] to obtain the following bifactor approximation:

Theorem 1. [4] There is an algorithm that, given an instance of Capacitated Location Routing and a number $\gamma \in (0, 1)$, computes in polynomial time a solution fulfilling conditions 1 to 3, with cost at most $(2 + 6/(1 - \gamma))\,\mathrm{OPT}$ and a maximum facility load of $(3/2 + 1/\gamma)B$.

It has further been observed that the known lower bounds based on Facility Location/Minimum Spanning Tree are not sufficient to derive constant factor approximations without exceeding the capacities [4].

References

1 R. Ravi R. and Sinha A. *Approximation algorithms for problems combining facility location and network design.* Operations Research, 54:73-81, 2006.

2 Harks T., König F.G., and Matuschke J. *Approximation algorithms for capacitated location routing.* Transportation Science, 47:3-22, 2013.

3 Lenstra J.K., Shmoys D.B., and Tardos E. *Approximation algorithms for scheduling unrelated parallel machines.* Mathematical Programming, 46:259-271, 1990.

4 Demleitner A. *Approximation Algorithms for Location Routing with Depot Capacities.* Master Thesis, Technische Universität München, 2019.

4.6 Scheduling on two types of machines

Bertrand Simon (Universität Bremen, DE)

We consider the problem of scheduling tasks with precedence constraints on several machines in order to minimize the makespan. The originality is that these machines are composed of two types of unrelated machines. Hence, this problem lies between the setting of identical and

[1] A tour T consists of a node set $V(T) \subseteq C \cup F$ and a permutation σ of $V(T)$. The length of T is $d(T) = \sum_{v \in V(T)} d(v, \sigma(v))$.

unrelated machines. Current results for the offline setting include a $3 + 2\sqrt{2}$-approximation (~ 5.8), and a conditional lower bound of 3 on the approximation ratio, assuming a variant of the unique games conjecture. Hence, there is still a gap between these bounds.

References

1 Fagnon V., Kacem I., Lucarelli G., and Simon B. *Scheduling on Hybrid Platforms: Improved Approximability Window.* LATIN, 2020.

2 Kedad-Sidhoum S., Monna F., and Trystram d. *Scheduling tasks with precedence constraints on hybrid multi-core machines.* IPDPS Workshop, 2015.

Participants

- Antonios Antoniadis
  MPI für Informatik –
  Saarbrücken, DE
- Yossi Azar
  Tel Aviv University, IL
- Etienne Bamas
  EPFL – Lausanne, CH
- Sanjoy K. Baruah
  Washington Univerity in
  St. Louis, US
- Shuchi Chawla
  University of Wisconsin –
  Madison, US
- Christian Coester
  CWI – Amsterdam, NL
- Sami Davies
  University of Washington –
  Seattle, US
- Christoph Dürr
  UPMC – Paris, FR
- Franziska Eberle
  Universität Bremen, DE
- Thomas Erlebach
  University of Leicester, GB
- Naveen Garg
  Indian Institute of Technology –
  New Delhi, IN
- Tobias Harks
  Univeristät Augsburg, DE
- Ruben Hoeksma
  University of Twente –
  Enschede, NL
- Sungjin Im
  University of California –
  Merced, US
- Sven Jäger
  TU Berlin, DE
- Xinrui Jia
  EPFL – Lausanne, CH
- Thomas Kesselheim
  Universität Bonn, DE
- Samir Khuller
  Northwestern University –
  Evanston, US
- Max Klimm
  HU Berlin, DE
- Peter Kling
  Universität Hamburg, DE
- Amit Kumar
  Indian Inst. of Technology –
  New Dehli, IN
- Marilena Leichter
  TU München, DE
- Jan Karel Lenstra
  CWI – Amsterdam, NL
- Alberto Marchetti-Spaccamela
  Sapienza University of Rome, IT
- Sebastien Martin
  LYFT – New York, US
- Jannik Matuschke
  KU Leuven, BE
- Nicole Megow
  Universität Bremen, DE
- Rolf H. Möhring
  TU Berlin, DE
- Sarah Morell
  TU Berlin, DE
- Benjamin J. Moseley
  Carnegie Mellon University –
  Pittsburgh, US
- Kamesh Munagala
  Duke University – Durham, US
- Viswanath Nagarajan
  University of Michigan –
  Ann Arbor, US
- Seffi Naor
  Technion – Haifa, IL
- Neil Olver
  London School of Economics and
  Political Science, GB
- Britta Peis
  RWTH Aachen, DE
- Kirk Pruhs
  University of Pittsburgh, US
- Jens Quedenfeld
  TU München, DE
- Shijin Rajakrishnan
  Cornell University – Ithaca, US
- Lars Rohwedder
  EPFL – Lausanne, CH
- Thomas Rothvoss
  University of Washington –
  Seattle, US
- Guido Schäfer
  CWI – Amsterdam, NL
- Kevin Schewior
  Universität Köln, DE
- Jiri Sgall
  Charles University – Prague, CZ
- David Shmoys
  Cornell University, US
- Bertrand Simon
  Universität Bremen, DE
- René Sitters
  VU University of Amsterdam, NL
- Martin Skutella
  TU Berlin, DE
- Clifford Stein
  Columbia University, US
- Leen Stougie
  CWI – Amsterdam, NL
- Ola Svensson
  EPFL – Lausanne, CH
- Éva Tardos
  Cornell University – Ithaca, US
- Vera Traub
  Universität Bonn, DE
- Marc Uetz
  University of Twente –
  Enschede, NL
- Rob van Stee
  Universität Siegen, DE
- Laura Vargas Koch
  RWTH Aachen, DE
- Victor Verdugo
  London School of Economics, GB
- Jose Verschae
  Pontifical Catholic University of
  Chile – Santiago, CL
- Tjark Vredeveld
  Maastricht University, NL
- Andreas Wiese
  Universidad de Chile –
  Santiago, CL

Report from Dagstuhl Seminar 20091

SE4ML – Software Engineering for AI-ML-based Systems

Edited by

Kristian Kersting[1], Miryung Kim[2], Guy Van den Broeck[3], and Thomas Zimmermann[4]

1 **TU Darmstadt, DE,** kersting@cs.tu-darmstadt.de
2 **UCLA, US,** miryung@cs.ucla.edu
3 **UCLA, US,** guyvdb@cs.ucla.edu
4 **Microsoft Corporation – Redmond, US,** tzimmer@microsoft.com

Abstract

Multiple research disciplines, from cognitive sciences to biology, finance, physics, and the social sciences, as well as many companies, believe that data-driven and intelligent solutions are necessary. Unfortunately, current artificial intelligence (AI) and machine learning (ML) technologies are not sufficiently democratized – building complex AI and ML systems requires deep expertise in computer science and extensive programming skills to work with various machine reasoning and learning techniques at a rather low level of abstraction. It also requires extensive trial and error exploration for model selection, data cleaning, feature selection, and parameter tuning. Moreover, there is a lack of theoretical understanding that could be used to abstract away these subtleties. Conventional programming languages and software engineering paradigms have also not been designed to address challenges faced by AI and ML practitioners. In 2016, companies invested \$26–39 billion in AI and McKinsey predicts that investments will be growing over the next few years. Any AI/ML-based systems will need to be built, tested, and maintained, yet there is a lack of established engineering practices in industry for such systems because they are fundamentally different from traditional software systems.

This Dagstuhl Seminar brought together two rather disjoint communities together, software engineering and programming languages (PL/SE) and artificial intelligence and machine learning (AI-ML) to discuss open problems on how to improve the productivity of data scientists, software engineers, and AI-ML practitioners in industry.

Seminar February 23–28, 2020 – http://www.dagstuhl.de/20091
2012 ACM Subject Classification Software and its engineering, Computing methodologies → Artificial intelligence, Computing methodologies → Machine learning
Keywords and phrases correctness / explainability / traceability / fairness for ml, data scientist productivity, debugging/ testing / verification for ml systems
Digital Object Identifier 10.4230/DagRep.10.2.76

1 Executive Summary

Kristian Kersting (TU Darmstadt, DE)
Miryung Kim (UCLA, US)
Guy Van den Broeck (UCLA, US)
Thomas Zimmermann (Microsoft Corporation – Redmond, US)

Any AI- and ML-based systems will need to be built, tested, and maintained, yet there is a lack of established engineering practices in industry for such systems because they are fundamentally different from traditional software systems. Building such systems requires extensive trial and error exploration for model selection, data cleaning, feature selection, and parameter tuning. Moreover, there is a lack of theoretical understanding that could be used to abstract away these subtleties. Conventional programming languages and software engineering paradigms have also not been designed to address challenges faced by AI and ML practitioners. This seminar brainstormed ideas for developing a new suite of ML-relevant software development tools such as debuggers, testers and verification tools that increase developer productivity in building complex AI systems. It also discussed new innovative AI and ML abstractions that improve programmability in designing intelligent systems.

The seminar brought together a diverse set of attendees, primarily coming from two distinct communities: software engineering and programming languages vs. AI and machine learning. Even within each community, we had attendees with various backgrounds and a different emphasis in their research. For example, within software engineering the profile of our attendees ranged from pure programming languages, development methodologies, to automated testing. Within, AI, this seminar brought together people on the side of classical AI, as well as leading experts on applied machine learning, machine learning systems, and many more. We also had several attendees coming from adjacent fields, for example attendees whose concerns are closer to human-computer interaction, as well as representatives from industry. For these reasons, the first two days of the seminar were devoted to bringing all attendees up to speed with the perspective that each other field takes on the problem of developing, maintaining, and testing AI/ML systems.

On the first day of the seminar, Ahmed Hassan and Tim Menzies represented the field of software engineering. Their talks laid the foundation for a lot of subsequent discussion by presenting some key definitions in software engineering for machine learning (SE4ML), identifying areas where there is a synergy between the fields, informing the seminar about their experiences dealing with industry partners, and listing some important open problems. Sameer Singh and Christopher Ré took care of the first day's introduction to machine learning. Christopher Ré described recent efforts in building machine learning systems to help maintain AI/ML systems, specifically for managing training data, and monitoring a deployed system to ensure it keeps performing adequately. Sameer Singh's talk focused on bug finding, and debugging machine learning systems, either by inspecting black-box explanations, generating realistic adversarial examples in natural language processing (NLP), and doing behavioral testing of NLP models to make them more robust.

The second day of the seminar continued to introduce the attendees to some prominent approaches for tackling the SE4ML problem. Elena Glassman presented her work at the intersection of human-computer interaction and software engineering, while Jie Zhang gave an overview of software testing for ML, based on her recent survey of the field. Significant attention during the seminar was spent on the problem of deploying machine learning models in environments that change over time, where the behavior of the AI/ML system diverges from the intended behavior when the model was first developed. For example, such issues were discussed by Barbara Hammer in her talk on machine learning in non-stationary environments. Isabel Valera introduced the seminar to another important consideration when developing AI/ML-based systems: interpretability and algorithmic fairness. Andrea Passerini's talk was aimed at explaining some of the basic principles of machine learning for a non-machine learning audience; for example generalization, regularization, and overfitting, as well as some recent trands in combining learning with symbolic reasoning.

The remainder of the seminar was centered around various breakout sessions and working groups, including sessions on (1) Specifications and Requirements, (2) Debugging and Testing, (3) Model Evolution and Management, and (4) Knowledge Transfer and Education. There were extended discussions on the question “what is a bug?” in an AI/ML setting, what is a taxonomy of such bugs, and can we list real-world examples of such bugs happening in practice. Interleaved with these working groups, there were several demand-driven talks, designed to answer questions that came up during the discussions. For example, Steven Holtzen and Parisa Kordjamshidi introduced the seminar to efforts in the AI community to build higher-level languages for machine learning, in particular probabilistic programming and declaritive learning-based programming. Christian Kästner shared his insights from teaching software engineering for AI/ML-based systems using realistic case studies. Molham Aref gave his unique view on developing such systems from industry, which was a tremendously valuable perspective to include in these discussions.

Overall, this seminar produced numerous new insights into how complex AI-ML systems are designed, debugged, and tested. It was able to build important scientific bridges between otherwise disparate fields, and has spurred collaborations and follow-up work.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 Machine Learning in non-stationary environments

Barbara Hammer (Universität Bielefeld, DE)

Joint work of Michael Biehl, Viktor Losing, Benjamin Paassen, Alexander Schulz, Heiko Wersing

One of the main assumptions of classical machine learning is that data are generated by a stationary concept. This, however, is violated in practical applications e.g. in the context of life long learning, for the task of system personalisation, or whenever sensor degradation or non-stationary environments cause a fundamental change of the observed signals. Within the talk, we will give an overview about recent developments in the field of learning with concept drift, and we will address two particular challenges in more detail: (1) How to cope with a fundamental change of the data representation which is caused e.g. by a misplacement or exchange of sensors? (2) How to deal with heterogeneous concept drift, i.e. mixed rapid or smooth, virtual or real drift, e.g. caused by a real-life non-stationary environment? We will present novel intuitive distance-based classification approaches which can tackle such settings by means of suitable metric learning and brain-inspired adaptive memory concepts, respectively, and we will demonstrate their performance in different application domains ranging from computer vision to the control of protheses.

References

1 Viktor Losing, Taizo Yoshikawa, Martina Hasenjäger, Barbara Hammer, Heiko Wersing: Personalized Online Learning of Whole-Body Motion Classes using Multiple Inertial Measurement Units. ICRA 2019: 9530-9536
2 Michiel Straat, Fthi Abadi, Christina Göpfert, Barbara Hammer, Michael Biehl: Statistical Mechanics of On-Line Learning Under Concept Drift. Entropy 20(10): 775 (2018)
3 Viktor Losing, Barbara Hammer, Heiko Wersing: Incremental on-line learning: A review and comparison of state of the art algorithms. Neurocomputing 275: 1261-1274 (2018)
4 Benjamin Paaßen, Alexander Schulz, Janne Hahne, Barbara Hammer: Expectation maximization transfer learning and its application for bionic hand prostheses. Neurocomputing 298: 122-133 (2018)
5 Viktor Losing, Barbara Hammer, Heiko Wersing: Tackling heterogeneous concept drift with the Self-Adjusting Memory (SAM). Knowl. Inf. Syst. 54(1): 171-201 (2018)
6 Viktor Losing, Barbara Hammer, Heiko Wersing: Self-Adjusting Memory: How to Deal with Diverse Drift Types. IJCAI 2017: 4899-4903
7 Viktor Losing, Barbara Hammer, Heiko Wersing: Personalized maneuver prediction at intersections. ITSC 2017: 1-6

3.2 Data Driven Decision Making for the Development of Trustworthy Software

Ahmed E. Hassan (Queen's University – Kingston, CA)

Software systems produce an enormous amount of rich data while being used (e.g., crashes, logs, telemetry data, and user reviews) and while being developed (e.g., historical code changes, test results, and feature requests). Leveraging such rich data through machine learning (ML), we can deliver better software in a cost-effective manner.

In this talk, I share my team's experience working closely with industrial partners over the past decade to address software development and operation (e.g., AIOps) challenges using ML. Then I discuss essential technical and non-technical goals to ensure the long-term successful integration of such ML solutions into daily practice.

3.3 Probabilistic Programming

Steven Holtzen (UCLA, US)

This talk provides a gentle introduction to probabilistic modeling and probabilistic programs. First, we ask what is a probabilistic program and how can they be used to solve problems? After some motivating examples, we discuss challenges in automating probabilistic inference. We highlight several example probabilistic programming languages and their diverse approaches to probabilistic inference, including (1) Stan [1], (2) Problog [2], (3) Dice [3], and (4) Figaro [4]. We close with a discussion of existing systems and prospects for integrating probabilistic programs and software engineering.

References

1 Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2016. *Stan: A probabilistic programming language.* Journal of Statistical Software (2016)

2 Daan Fierens, Guy Van den Broeck, Joris Renkens, Dimitar Shterionov, Bernd Gutmann, Ingo Thon, Gerda Janssens, and Luc De Raedt. 2013. *Inference and learning in probabilistic logic programs using weighted Boolean formulas.* J. Theory and Practice of Logic Programming 15(3) (2013), 358 – 401.

3 Steven Holtzen, Guy Van den Broeck, and Todd Millstein. 2020. *Dice: Compiling Discrete Probabilistic Programs for Scalable Inference.* arXiv preprint arXiv:2005.09089.

4 Avi Pfeffer. 2009. *Figaro: An object-oriented probabilistic programming language.* Charles River Analytics Technical Report 137 (2009).

3.4 Declarative Learning-Based Programming as an Interface to AI Systems

Parisa Kordjamshidi (Michigan State University – East Lansing, US)

Joint work of Parisa Kordjamshidi, Kristian Kersting, Dan Roth
Main reference Parisa Kordjamshidi, Dan Roth, Kristian Kersting: "Declarative Learning-Based Programming as an Interface to AI Systems", CoRR, Vol. abs/1906.07809, 2019.
URL https://arxiv.org/abs/1906.07809

Data-driven approaches are becoming more common as problem-solving techniques in many areas of research and industry. In most cases, machine learning models are the key component of these solutions, but a solution involves multiple such models, along with significant levels of reasoning with the models' output and input. Current technologies do not make such techniques easy to use for application experts who are not fluent in machine learning nor for machine learning experts who aim at testing ideas and models on real-world data in the context of the overall AI system. We review key efforts made by various AI communities to provide languages for high-level abstractions over learning and reasoning techniques needed for designing complex AI systems. We classify the existing frameworks based on the type of techniques and the data and knowledge representations they use, provide a comparative study of the way they address the challenges of programming real-world applications, and highlight some shortcomings and future directions.

3.5 Teaching Software Engineering for AI-enabled Systems

Christian Kästner (Carnegie Mellon University – Pittsburgh, US)

Main reference Christian Kästner, Eunsuk Kang: "Teaching Software Engineering for AI-Enabled Systems", 2020.
URL https://arxiv.org/abs/2001.06691.9

Software engineers have significant experience to offer when building intelligence systems, drawing on decades of methods for building systems that scale and are robust, even when built on unreliable components. Systems with AI/ML components raise new challenges and require careful engineering, for which we designed a new course. We specifically go beyond traditional ML courses that teach modeling techniques under artificial conditions and focus on realism with large and changing datasets, robust and evolvable infrastructures and requirements engineering that considers also ethics and fairness. We share all course material

- Software Engineering for AI-Enabled Systems (SE4AI)
  https://ckaestne.github.io/seai/
- Software Engineering for AI/ML – An Annotated Bibliography
  https://github.com/ckaestne/seaibib

3.6 SE for (AI+SE)

Tim Menzies (North Carolina State University – Raleigh, US)

What should this community tell the world abut SE and AI? What are our "seven deadly sins" and our "dozen" best practices?

To answer these questions, I offer (tiny) summaries of SE and AI practice. The focus here will be "what are the surprises?", i.e., what are the *new* things we know *now* that we didn't know before. For example

- "Programmers" do much more than programming. And in fact, social factors between programming can be just as predictive for bugs as any programming language feature.
- Some (not all) SE data is inherently low dimensional and we can exploit that great benefit.

For more on this talk, see http://tiny.cc/se4ml

3.7 Some Machine Learning Basics + Random Stuff

Andrea Passerini (University of Trento, IT)

I will give a quick overview of the basic concepts used in machine learning from generalisation to regularized loss minimizations to model selection. I will quickly present how to use the scikit learn framework to train multiple classifiers and give a bird's eye view of deep learning. I will end up presenting some work of mine focused on the combination of learning and constraints.

3.8 Experiences Building & Maintaining Software 2.0 Systems.

Christopher Ré (Stanford University, US)

This talk describes our group's recent working building and maintaining a new breed of ML software systems. The talk focusses on how engineer time is spent in building and maintaining these systems. Two main example systems were discussed.

1. Snorkel. A system to make creating and maintaining training sets a 1st class problem in both software and statistical theory.
2. Overton A system built at Apple that focused engineer time on maintaining supervision and monitoring its output quality – not more building.

3.9 Testing and Finding Bugs in NLP Models

Sameer Singh (University of California – Irvine, US)

Current evaluation of NLP systems, and much of ML, consists of measuring accuracy on held-out instances. Since held-out instances are gathered using similar annotation process as the training data, they include the same biases, providing "shortcuts" to NLP models. Further, single aggregate metric hides the actual strengths and weaknesses of the model, making it difficult to focus engineering and research efforts.

In this talk, I presented a few approaches we are exploring to perform a more thorough evaluation of NLP systems.

1. I will introduce our work on *generating black-box explanations* for ML models (LIME and Anchors) and their use in finding bugs.
2. I will describe automatic techniques for perturbing instances to identify shortcuts via *semantic adversarial examples*.
3. I will propose novel ML paradigms that introduce "testing for ML", in particular *Checklist* for creating behavioral tests for NLP.

The talk will be grounded in latest NLP benchmarks such as QA, sentiment analysis, and textual entailment, on SOTA models like BERT.

3.10 ML for Consequential Decision Making

Isabel Valera (MPI für Intelligente Systeme – Tübingen, DE)

This talk provided a brief overview of fairness and interpretability in ML, painting out the main challenges in the topic. Then, I introduce an example of how formal verification approaches can help explainable ML by providing with a model and similarity agnostic, as well as modular framework to generate (nearest) counterfactual explanations for the outcomes of algorithmic decision making systems. This example was later extended with some existing work on software engineering for adversarial robustness in ML. We close the presentation opening up questions on how software engineering may be helpful to define, test, and verify specification on the ethics of ML

- Some references on Counterfactual explanations:
  https://arxiv.org/pdf/1905.11190.pdf
  https://arxiv.org/abs/2002.06278
- Some work on fairness:
  http://jmlr.org/papers/v20/18-262.html
  https://arxiv.org/abs/1902.02979

3.11 Software Testing for Machine Learning

Jie Zhang (University College London, GB)

Main reference J. M. Zhang, M. Harman, L. Ma, Y. Liu: "Machine Learning Testing: Survey, Landscapes and Horizons", IEEE Transactions on Software Engineering, pp. 1–1, 2020.
URL https://doi.org/10.1109/TSE.2019.2962027

Machine learning systems are a type of software. This talk builds the connection between software testing and machine learning.

I first gave a brief introduction on software testing. Software testing aims to evaluate a software to check whether its behaviors meet the requirements. I introduced the properties of interest, the testing component, the testing workflow, and some key techniques in automated software testing.

Based on traditional software testing, I introduced machine learning testing (MLT). MLT detects the imperfections in machine learning systems that violate the expectation. The properties of interest may include correctness, fairness, privacy, security, interpretability. Different from traditional software testing, MLT bugs may exist in the data, learning programs, or frameworks. Many traditional testing techniques can be adopted in MLT.

I gave an overview of the related work in MLT so far. The details of the related work can be found in our survey: *Machine Learning Testing: Survey, Landscapes, and Horizons* (TSE 2020)

The last part of my talk is about my two recent practices in improving ML systems.

- *Perturbation Validation (PV)* is a compliment for out-of-sample validation in model validation. It does not use validation or test but checks whether the learner detects a small ratio of incorrect labels in training data.

- *Black-box repair* fixes machine translation problems without model retraining, so it is automatic, fast, light-weight, and can target and repair specific cases without touching other well-formed translations.

4 Working groups

4.1 Agile Development of AI/ML-based Systems

Andreas Metzger (Universität Duisburg – Essen, DE), Christian Kästner (Carnegie Mellon University – Pittsburgh, US), and Daniel Speicher (Universität Bonn, DE)

This breakout working session aimed at identifying how typical management and engineering practices of agile methods may be affected when developing AI/ML-based systems. To this end, typical practices of XP (eXtreme Programming [1]) were used to serve to structure the discussions. The main outcomes of this session were open questions that may provide an opportunity for further investigation, such as empirical studies.

The practice of the *planning game* in XP allows customers and developers to steer the work towards the most useful system the team can deliver. Functionality or quality increments are described, often estimated for the required implementation effort, and finally selected. Customers contribute their knowledge about the business value of increments. Developers contribute their knowledge about technical complexity and risks. Questions regarding the planning game included: (1) Is effort planning for AI/ML components more challenging and less precise (e.g., since creating an AI/ML model may be more explorative and experimental)? (2) How to assess the value contribution of AI/ML components? (3) Can sufficiently small work items (i.e., user stories) be defined?

The practice of *pair programming* (worth its own book [2]) has several goals, such as knowledge diffusion and skill transfer within the team. This keeps the team in the position to evolve every part of the system even when a member of the team leaves. The practice of *collective code ownership* allows all team members to (carefully) change any part of the system. Questions regarding pair programming and collective code ownership were: (1) How can AI/ML experts and software engineers work together? (2) Given joint teams of AI/ML experts and software engineers, what may happen if either may change a machine learning model and program code? (3) Do we need these practices for AI/ML components at all (e.g., concerns such as technical debt may be addressed via AutoML etc.)?

The practice of *simple design* encourages developers to stay with simple solutions. Simpler solutions lead to faster results and allow earlier customer feedback. Unnecessary technological complexity may burden future change and development. "Simplicity" here is not an absolute term, but relative to the team's knowledge and experience. The central question regarding simple design was: How to make a trade-off between deep learning and "shallow" learning? While deep learning may generally lead to less interpretable and explainable AI/ML models than "shallow" learning, deep learning requires less feature engineering and thus requires less engineering effort.

The practice of *refactoring* has the goal to keep the design simple and to maintain code quality (such as maintainability, changeability, understandability). To guide refactoring, developers have described refactoring opportunities, often called "code smells". Questions regarding refactoring included: (1) How can AI/ML model quality be defined in the first place? (2) How to refactor towards good AI/ML model quality? (3) As AI/ML models are generated and not hand-crafted, is refactoring needed at all?

The practice of *test-driven development* has the goal to establish a solid base of automated tests and to guide development. Also, automated tests safeguard existing functionality during refactoring and functionality addition. The key question regarding test-driven development was: How to define feasible test cases up-front (and not just non-functional constraints on the output of the AI/ML model)? We realized that the answer to this question was very much tied to the problem of how to specify AI/ML-based systems and whether machine learning may be requirements engineering – a topic that was discussed throughout the seminar [3].

References

1 Kent Beck. *Extreme Programming Explained: Embrace Change*. Second Edition. Addison-Wesley, Reading, MA, 2005

2 Laurie Williams, Robert Kessler. *Pair programming illuminated.* Addison-Wesley Longman Publishing Co., Inc., 2002.

3 Christian Kästner. "Machine Learning is Requirements Engineering – On the Role of Bugs, Verification, and Validation in Machine Learning", Medium post, Accessed April 25, 2020. https://medium.com/analytics-vidhya/machine-learning-is-requirements-engineering-8957aee55ef4

Participants

- Hadil Abukwaik
ABB – Ladenburg, DE
- Molham Aref
relationalAI – Berkeley, US
- Earl T. Barr
University College London, GB
- Houssem Ben Braiek
Polytechnique Montréal, CA
- Pavol Bielik
ETH Zürich, CH
- Carsten Binnig
TU Darmstadt, DE
- Luc De Raedt
KU Leuven, BE
- Rob DeLine
Microsoft Corporation – Redmond, US
- Joachim Giesen
Universität Jena, DE
- Elena Leah Glassman
Harvard University – Cambridge, US
- Nikolas Göbel
RelationalAI – Zürich, CH
- Jin L.C. Guo
McGill University – Montréal, CA
- Barbara Hammer
Universität Bielefeld, DE
- Fabrice Harel-Canada
UCLA, US
- Ahmed E. Hassan
Queen's University – Kingston, CA
- Steven Holtzen
UCLA, US
- Christian Kästner
Carnegie Mellon University – Pittsburgh, US
- Kristian Kersting
TU Darmstadt, DE
- Miryung Kim
UCLA, US
- Angelika Kimmig
Cardiff University, GB
- Parisa Kordjamshidi
Michigan State University – East Lansing, US
- Vu Le
Microsoft Corporation – Redmond, US
- Rupak Majumdar
MPI-SWS – Kaiserslautern, DE
- Tim Menzies
North Carolina State University – Raleigh, US
- Andreas Metzger
Universität Duisburg – Essen, DE
- Mira Mezini
TU Darmstadt, DE
- Alejandro Molina
TU Darmstadt, DE
- Sandeep Neema
DARPA – Arlington, US
- Siegfried Nijssen
UC Louvain, BE
- Andrea Passerini
University of Trento, IT
- Michael Pradel
Universität Stuttgart, DE
- Christopher Ré
Stanford University, US
- Sameer Singh
University of California – Irvine, US
- Daniel Speicher
Universität Bonn, DE
- Isabel Valera
MPI für Intelligente Systeme – Tübingen, DE
- Guy Van den Broeck
UCLA, US
- Antonio Vergari
UCLA, US
- Laurie Williams
North Carolina State University – Raleigh, US
- Ce Zhang
ETH Zürich, CH
- Jie Zhang
University College London, GB
- Tianyi Zhang
Harvard University – Cambridge, US
- Xiangyu Zhang
Purdue University – West Lafayette, US
- Thomas Zimmermann
Microsoft Corporation – Redmond, US