DAGSTUHL
REPORTS

Volume 4, Issue 10, October 2014

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
February, 2015

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.4.10.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 14411

Constraints, Optimization and Data

Edited by

Luc De Raedt[1], Siegfried Nijssen[1,2], Barry O'Sullivan[3], and Michele Sebag[4]

1 **KU Leuven, BE,** `firstname.lastname@cs.kuleuven.be`
2 **Universiteit Leiden, NL,** `s.nijssen@liacs.leidenuniv.nl`
3 **University College Cork, IE,** `b.osullivan@cs.ucc.ie`
4 **Université Paris Sud, FR,** `michele.sebag@lri.fr`

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 14411 "Constraints, Optimization and Data". Constraint programming and optimization have recently received considerable attention from the fields of machine learning and data mining; similarly, machine learning and data mining have received considerable attention from the fields of constraint programming and optimization. The goal of the seminar was to showcase recent progress in these different areas, with the objective of working towards a common basis of understanding, which should help to facilitate future synergies.

Seminar October 6–10, 2014 – http://www.dagstuhl.de/14411
1998 ACM Subject Classification H.2.8 Database applications. F.4.1. Mathematical logic
Keywords and phrases Data mining, constraint programming, machine learning
Digital Object Identifier 10.4230/DagRep.4.10.1

1 Executive Summary

Luc De Raedt
Siegfried Nijssen
Barry O'Sullivan
Michele Sebag

Constraint programming and optimization (CPO) have recently received considerable attention from the fields of machine learning and data mining (MLDM). On the one hand, the hypotheses and patterns that one seeks to discover in MLDM can be specified in terms of constraints (e. g. labels in the case of supervised learning, preferences in the case of learning to rank, must-link and cannot-link in the case of unsupervised learning, coverage and lift in the case of data mining). On the other hand, powerful constraint programming solvers have been developed. If MLDM users express their requirements in terms of constraints they can delegate the MLDM process to such highly efficient solvers.

Conversely, CPO can benefit from integrating learning and mining functionalities in a number of ways. For example, formulating a real-world problem in terms of constraints requires significant expertise in the problem domain. Also, selecting the most appropriate constraints, in terms of constraint solving efficiency, requires considerable expertise in the CPO domain. In other words, experience plays a major role in successfully applying CPO technology.

Constraints, Optimization and Data, *Dagstuhl Reports*, Vol. 4, Issue 10, pp. 1–31
Editors: Luc De Raedt, Siegfried Nijssen, Barry O'Sullivan, and Michele Sebag
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In addition, both CPO and MLDM share a common challenge associated with tuning their respective methods, specifically determining the best parameters to chose for an algorithm depending on the task at hand. A typical performance metric in machine learning is the predictive accuracy of a hypotheses, while in CPO it might be search cost or solution quality.

This seminar built upon the 2011 *Constraint Programming meets Machine Learning and Data Mining*[1] and the 2014 *Preference learning*[2] seminars. Its goal was to identify the key challenges and opportunities at the crossroads of CPO and MLDM. The interests of the participants included the following:

- Problem formulation and modelling: constraint-based modelling; preference formalisms; loss functions in ML; modelling and exploiting background knowledge; structured properties (e. g. preserving spatio-temporal structures).
- Improvement of algorithms / platforms in the areas of algorithm selection, algorithm configuration, and/or algorithm scheduling, particularly with respect to parallel execution.
- Specification and reasoning about goals and optimization criteria: modelling preferences and integrating with human expertise (exploiting the "human in the loop") to converge on high quality outcomes.
- Additional functionalities such as the use of visualization and explanation.
- Algorithmic scalability.
- Approximate reasoning, reasoning under uncertainty, and incorporating probability.

The seminar was organized into seven sessions: frameworks and languages; algorithm configuration; constraints in pattern mining; learning constraints; machine learning with constraints; applications; and demonstrations. The demonstrations presented at the seminar were by:

- Guido Tack – MiniZinc (see http://mininzinc.org);
- Joaquin Vanschoren – OpenML (see http://openml.org);
- Tias Guns – MiningZinc (see http://dtai.cs.kuleuven.be/CP4IM/miningzinc);
- Bruno Crémilleux – software for the calculation of Sky Pattern Cubes;
- Marc Denecker – IDP (see http://dtai.cs.kuleuven.be/krr/software/idp);
- Holger Hoos – algorithm selection and portfolio software;
- Luc De Raedt – ProbLog (see http://dtai.cs.kuleuven.be/problog/).

The seminar also had five working groups on:

- Declarative Languages for Machine Learning and Data Mining;
- Learning and Optimization with the Human in the Loop;
- Meta-Algorithmic Techniques;
- Big Data;
- Towards Killer Applications.

1 http://www.dagstuhl.de/11201
2 http://www.dagstuhl.de/14101

2 Table of Contents

3 Overview of Talks

3.1 The road to declarative data analysis is paved with constraints

Hendrik Blockeel (KU Leuven, BE)

Data analysis is declarative when the user can just formulate the question to be answered, rather than the method for solving that question. This talk presented arguments and illustrations for the following claims: (1) Declarative approaches will make data analysis easier, more flexible, more efficient and more correct. (2) Constraint-based reasoning will play a major part in reaching this objective.

3.2 Constraint-based mining: preliminary results on deriving constraints from experts models

Jean-François Boulicaut (INSA – Lyon, FR)

Joint work of Boulicaut, Jean-François; Flouvat, Fréderic; Selmaoui, Nazha; Sanhes, Jérémy
URL http://liris.cnrs.fr/~jboulica/

First, we discuss the basics of pattern domain design (specifying pattern languages over specific types of data, introducing measures and more or less related primitive constraints and the way to combine them as well). In fact, thanks to various recent work in our team (e. g., mining preserved cliques in relational dynamic graphs [1], gradual pattern mining in attributed graphs [2], mining sets of cohesive nodes in dynamic attribute graphs [3]), we can easily refer to the different roles of primitive constraints when considering one inductive query: some express pattern semantics, others are dedicated to pattern objective interestingness while the remaining ones specify subjective interestingness issues. We can then consider both the declarative and the computational views on the specified mining tasks. It is important to address the relationships of primitive constraint properties w.r.t. typical enumeration strategies over the pattern language. Notice that the terminology about constraint-based data mining and inductive database topics can be found in the following books [4, 5, 6].

In a second step, we report about recent results that we obtained thanks to a cooperation between INSA de Lyon and the University of New Caledonia. We consider the context where expert models are available as multivariate real functions. Then, we study how to derive a new type of primitive constraint that can exploit such a model by specifying that the model may predict a value higher than a user-defined threshold for the interesting patterns. As a concrete example, we introduce a case study on soil erosion in New Caledonia. We set up a simple item set mining context where transactions correspond to areas (more precisely pixels in satellite images) and items corresponds to various properties of the areas. Using the new constraint derived from models that evaluate an erosion risk, we can enforce the data mining tasks to focus on itemsets that are associated to areas for which the model predict a high enough risk. Studying the properties of the new primitive constraint and finding pruning rules that can be easily combined with the other pruning rules for item set mining is the main technical contribution. Perspectives are obvious like deriving new primitive constraints from the same kind of model, looking for new king of models, looking for more sophisticated data mining tasks (than just item set mining). This preliminary work has been published in [7].

References

1 Loïc Cerf, Bao Nhan Nguyen Tran, Jean-Francois Boulicaut. Mining constrained cross-graph cliques in dynamic networks. In: Inductive Databases and Constraint-Based Data Mining, S. Dzeroski, B. Goethals, and P. Panov (Eds), pp. 199–228, November 2010, Springer.

2 Adriana Prado, Marc Plantevit, Céline Robardet, Jean-Francois Boulicaut. Mining graph topological patterns: Finding co-variations among vertex descriptors. IEEE Transactions on Data and Knowledge Engineering 25(9):2090–2104, 2013, IEEE Computer Press.

3 Elise Desmier, Marc Plantevit, Céline Robardet, Jean-Francois Boulicaut. Trend mining in dynamic attributed graphs. Proc. European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML PKDD 2013, Praha, Czech Republic, September 2013. H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezný (Eds.). Springer LNAI 8188, pp. 654–669.

4 Database support for Data Mining Applications – Discovering Knowledge with Inductive Queries LNCS 2682, R. Meo, P-L. Lanzi, and M. Klemettinen (Eds.), 2004. Springer.

5 Constraint-based Mining and Inductive Databases, J-F. Boulicaut, L. De Raedt, H. Mannila (Eds.) Springer LNCS 3848, 2005.

6 Inductive Databases and Constraint-Based Data Mining, S. Dzeroski, B. Goethals, and P. Panov (Eds), November 2010, Springer.

7 Frédéric Flouvat, Jérémy Sanhes, Claude Pasquier, Nazha Selmaoui-Folcher, Jean-Francois Boulicaut. Improving pattern discovery relevancy by deriving constraints from expert models. Proc. European Conference on Artificial Intelligence ECAI 2014. August 2014, Praha, Czech Republic, T. Schaub et al. (Eds), regular paper, pp. 327–332, IOS.

3.3 Data, learning and optimisation in home energy management

Ken Brown (University College Cork, IE)

Joint work of Brown, Ken; Manzano, Oscar; Murphy, Sean Og; Trabelsi, Walid; O'Toole, Liam; O'Sullivan, Barry

The Authentic project is developing an integrated hardware and software solution of home energy management, to provide personal tailored feedback and guidance to home users. Data on appliance use, occupancy, environmental conditions and energy consumption are gathered by sensors. We then learn models of occupant behaviour and house response, identifying areas of energy inefficiency. Finally, we develop optimisation models to guide occupants towards achieving their own energy goals. The models attempt to find the minimal change to observed behaviour which achieves the goals. The models are learned online through interaction with the occupants.

3.4 On Preference-based (soft) pattern sets

Bruno Crémilleux (Caen University, FR)

Joint work of Boizumault, Patrice; Crémilleux, Bruno; Loudni, Samir; Ugarte Rojas, Willy
Main reference W. Ugarte, P. Boizumault, S. Loudni, B. Crémilleux, "Computing skypattern cubes," in Proc. of the 21st European Conf. on Artificial Intelligence (ECAI'14), pp. 903–908, IOS Press, 2014.
URL http://dx.doi.org/10.3233/978-1-61499-419-0-903

In the last decade, the pattern mining community has witnessed a sharp shift from efficiency-based approaches to methods which can extract more meaningful patterns. Recently, new methods adapting results from multi criteria decision analyses such as Pareto efficiency, or skylines, have been studied. Within pattern mining, this novel line of research allows the easy expression of preferences according to a dominance relation on a set of measures and avoids the well-known threshold issue.

In this talk, we present the discovery of soft skyline patterns (or soft skypatterns) based on theoretical relationships with condensed representations of patterns and the dynamic constraint satisfaction problems framework. To avoid an apriori choice of measures, we propose to use the skypattern cube according to the set of measures. We show how to efficiently build the skypattern cube and provide a concise representation of the cube based on skypattern equivalence classes. Navigation trough the cube indicates differences and similarities between skypattern sets when a measure is added or removed, it highlights the role of the measures and helps to discover the most interesting skypattern sets. We set our work in the global picture of the discovery of pattern sets and k-pattern sets/n-ary patterns.

References

1 Willy Ugarte and Patrice Boizumault and Samir Loudni and Bruno Crémilleux and Alban Lepailleur Mining (soft-) skypatterns using dynamic CSP. *In* 11th Int. Conf. on Integration of Artificial Intelligence (AI) and Operations Research (OR) Techniques in Constraint Programming (CP) (CPAIOR 2014), 19–23 May, 2014, Cork, Ireland, pages 71–87.
2 Willy Ugarte and Patrice Boizumault and Samir Loudni and Bruno Crémilleux Computing skypattern cubes. *In* 21st European Conference on Artificial Intelligence (ECAI 2014), 18–22 August 2014, Prague, Czech Republic, pp. 903–908. IOS Press.
3 Willy Ugarte and Patrice Boizumault and Samir Loudni and Bruno Crémilleux Computing skypattern cubes using relaxation. *In* 26th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2014), 10–12 November 2014, Limassol, Cyprus.

3.5 GOBNILP: CIP for BN learning

James Cussens (University of York, GB)

URL http://www.cs.york.ac.uk/aig/sw/gobnilp

In this talk I will describe the Bayesian network learning system GOBNILP [1, 2] and describe more general issues concerned with using (constraint) integer programming (CIP) for machine learning tasks. A CIP allows one to describe a machine learning task declaratively and then have a solver solve the resulting problem. However, doing so *efficiently* is not so simple, and in our experience requires detailed study and considerable implementation effort. I will discuss the lessons we have learned in this work. Chief amongst these is that with a

CIP/MIP approach constructing a tight linear relaxation is crucial. To do this it is necessary to *avoid* compact representations of the problem.

References

1 James Cussens. Bayesian network learning with cutting planes. In Fabio G. Cozman and Avi Pfeffer, editors, *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pp. 153–160, Barcelona, 2011. AUAI Press.

2 Mark Barlett and James Cussens. Advances in Bayesian network learning using integer programming. In Ann Nicholson and Padhraic Smyth, editors, *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pp. 182–191, Bellevue, 2013. AUAI Press

3.6 Constrained Clustering by Constraint Programming

Thi-Bich-Hanh Dao (University of Orleans, FR)

Joint work of Dao, Thi-Bich-Hanh; Duong, Khanh-Chuong; Vrain, Christel
Main reference T.-B.-H. Dao, K.-C. Duong, C. Vrain, "Constrained Clustering by Constraint Programming," to appear.

Constrained Clustering has received much attention this last decade. It allows to make the clustering task either easier or more accurate by integrating user-constraints. Several kinds of constraints can be considered; they may be requirements on the clusters, as for instance their sizes or their diameters, or on pairs of instances the expert knows that they must be or cannot be in the same cluster (must-link or cannot-link constraints). Much work has focused on instance-based constraints and has adapted classical clustering methods to handle such kinds of constraints, but usually considering only one optimization criterion. Few works consider different kinds of constraints or integrate different optimization criteria. In [1] a SAT based framework for constrained clustering has been proposed, but it is limited to problems with two classes. A framework based on Integer Linear Programming has also been proposed in [2] but it is more suited to conceptual clustering. Another Integer Linear Programming framework has been proposed in [3], which considers must-link, cannot-link constraints and anti-monotone constraints, but with a single optimization criterion that is the minimum sum of squares.

In this talk, we present a framework based on Constraint Programming for Constrained Clustering [4, 5]. The framework is general and declarative, it allows to choose among several optimization criteria and to integrate different kinds of user-constraints. The optimization criteria proposed are minimizing the maximal diameter of the clusters, maximizing the minimal split between clusters or minimizing the within-cluster sum of dissimilarities. The framework integrates must-link or cannot-link constraints and all popular cluster level constraints. The model is flexible since it does not require to set the number of clusters beforehand, only a lower and an upper bound on the number of clusters have to be given. Relying on the dissimilarity between objects, the model can handle quantitative or qualitative datasets, as soon as such a measure can be defined.

The approach we propose can be easily embedded in a more general process for Constrained Clustering. Considering Data Mining as an iterative and interactive process composed of the classical steps of task formulation, data preparation, choice of a learning tool thus requiring to set parameters and validation of the results, a user can specify the task at hand including or not some constraints and decide to change the settings according to the results. He/she may decide to change the constraints, removing some constraints, adding or hardening other ones. The modularity and the declarativity of our model allow this easily. In this talk, we illustrate this point on a bi-criterion clustering problem. We show that our framework can be used to find the Pareto front of a bi-criterion maximizing split-minimizing diameter problem, under consideration of user-constraints.

References

1. I. Davidson, S.S. Ravi, L. Shamis. *A SAT-base Framework for Efficicient Constrained Clustering*. In Proceedings of the 10th SIAM International Conference on Data Mining, pp. 94–105, 2010.
2. M. Mueller, S. Kramer. *Integer Linear Programming Models for Constrained Clustering*. in Proceedings of the 13th International Conference on Discovery Science, pp. 159–173, 2010.
3. B. Babaki, T. Guns, S. Nijssen. *Constrained Clustering using Column Generation*. In Proceedings of the 11st International Conference on Integration of AI and OR Techniques in Constraint Programming, pp. 438–454, 2014.
4. T.-B.-H. Dao, K.-C. Duong, C. Vrain. *A Declarative Framework for Constrained Clustering*. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, pp. 419–434, 2013.
5. T.-B.-H. Dao, K.-C. Duong, C. Vrain. *Constrained Clustering by Constraint Programming*. Artificial Intelligence Journal, accepted with minor revisions.

3.7 Human in the Loop Learning with Constraints

Ian Davidson (University of California – Davis, US)

We show how constraints can be used for active and transfer learning to facilitate human in the loop learning. The constraints are used as a mechanism for the human to have a dialog with the algorithm. The human specifies guidance to the machine in the form of constraints and the algorithm asks questions of the human also in the form of constraints. We briefly overview applications in fMRI data analysis.

3.8 Languages for Mining and Learning

Luc De Raedt (KU Leuven, Belgium)

Applying machine learning and data mining to novel applications is cumbersome. This observation is the prime motivation for the interest in languages for learning and mining. This note provides a gentle introduction to three types of languages that support machine learning and data mining: inductive query languages, which extend database query languages with primitives for mining and learning, modelling languages, which allow to declaratively specify and solve mining and learning problems, and programming languages, that support the learning of functions and subroutines. It uses an example of each type of language to introduce the underlying ideas and puts them into a common perspective. This then forms the basis for a short analysis of the state-of-the-art.

References

1 Luc De Raedt. *Languages for Mining and Learning*. Proc. 29th AAAI Conference on Artificial Intelligence, Senior Member Track, 2015, in press.

3.9 Constraint solving with extensions of classical logic Applications to data mining

Marc Denecker (KU Leuven, BE)

I will present a motivation for the use of (extensions of) first order logic for constraint solving. I will situate this in the context of constraint programming and answer set programming. Examples will serve to present the paradigm and some useful language extensions. I will give a brief overview of the system IDP3 and discuss implementation techniques that have been used in it and results of experiments. I can discuss the relationship with Zinc, and illustrate applications of the system to some datamining applications.

3.10 Learning to Build Effective Constraint Models

Alan Frisch (University of York, GB)

Joint work of Frisch, Alan; Miguel, Ian; Jefferson, Chris; Agkun, Ozgur; Martinez-Hernandez, Bernadette
Main reference O. Akun, I. Miguel, C. Jefferson, A. M. Frisch, B. Hnich, "Extensible Automated Constraint Modelling," in Proc. of the 25th AAAI Conf. on Artificial Intelligence (AAAI'11), pp. 4–11, AAAI, 2011; pre-print available from author's webpage.
URL http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3687
URL http://www.cs.york.ac.uk/aig/constraints/AutoModel/conjure-aaai11.pdf

CONJURE is a rule-based system that can automatically generate a set of alternative constraint models for a problem expressed in the ESSENCE problem specification language. CONJURE currently has no highly effective method to select a good model from among a set of alternatives it generates.

After a brief introduction to ESSENCE and CONJURE, this talk argues that it is better to drive the learning of model selection not by features of the models but rather by features of the design decisions made in generating the model.

References

1 Alan M. Frisch, Warwick Harvey, Chris Jefferson, Bernadette Martinez Hernandez and Ian Miguel. ESSENCE: A Constraint Language for Specifying Combinatorial Problems. In: *Constraints*, 13(3), July 2008.
2 Ozgur Akgun, Ian Miguel, Chris Jefferson, Alan M Frisch and Brahim Hnich. Extensible Automated Constraint Modelling. In: *Proc. of the 25th AAAI Conference on Artificial Intelligence*, 2011.

3.11 Classes of constraints within a high level model of constraint optimization

Randy Goebel (University of Alberta, CA)

The traditional logic programming appeal of all computation falling within logical semantics extends to the architecture of constraint optimization, where constraint specifications of a model, an objective function, and method constraints can be uniformly specified under a single semantics. We sketch a simple uniform architecture which helps identify how incremental dynamic constraint optimization systems can be incrementally improved. The architecture shows the fit for where probability, continuous variables, hypothetical reasoning, and belief revision.

3.12 MiningZinc, a declarative framework for constraint-based mining

Tias Guns (KU Leuven, BE)

Joint work of Guns, Tias; Dries, Anton; Nijssen, Siegfried; Tack, Guido; De Raedt, Luc
Main reference T. Guns, A. Dries, G. Tack, S. Nijssen, L. De Raedt, "MiningZinc: A modeling language for constraint-based mining," in Proc. of the 23rd Int'l Joint Conf. on Artificial Intelligence (IJCAI'13), pp. 1365–1372, AAAI Press, 2013.
URL http://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6947
URL http://dtai.cs.kuleuven.be/CP4IM/miningzinc/

We presented MiningZinc, a declarative framework for constraint-based data mining. MiningZinc consists of two key components: a language component and an execution mechanism. The language allows for high-level and natural modeling of mining problems, such that MiningZinc models closely resemble the definitions found in the data mining literature. It is inspired by the Zinc family of languages and systems and supports user-defined constraints and functions.

The execution mechanism specifies how to compute solutions for the models. It ensures the solver independence of the language and supports both standard constraint solvers and specialized data mining systems. The high-level problem specification is first translated into a normalized constraint language (FlatZinc). Rewrite rules are then used to add redundant constraints or solve part of the problem specification using specialized algorithms or generic con- straint programming solvers. For one model, different execution strategies are automatically extracted that correspond to different sequences of algorithms to run.

In this way, MiningZinc combines high-level declarative modeling with the performance of state-of-the-art algorithms on well-known tasks. Furthermore, its language allows one to model constraint-based mining problems and varia- tions, and its execution mechanism can detect and re-use existing algorithms. We demonstrate this experimentally on a number of tasks.

References

1 T. Guns, A. Dries, G. Tack, S. Nijssen, L. De Raedt. MiningZinc: A modeling language for constraint-based mining. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press, 2013, pp. 1365–1372.

3.13 Analysing and Automatically Optimising the Empirical Scaling of Algorithm Performance

Holger H. Hoos (University of British Columbia – Vancouver, CA)

Joint work of Hoos, Holger H.; Stützle, Thomas; Styles, James

The scaling of running time with input size is of central importance in the analysis and design of algorithms. Traditionally, this is captured by the notion of time complexity and analysed mathematically. However, this analysis is often restricted to simplified or simplistic algorithms, ignores constants and considers only worst- case behaviour, or average-case behaviour under restrictive assumptions on the distribution of inputs. The same limitations apply to algorithm design or selection guided by theoretical time complexity.

In this talk, I first present a new approach for analysing the empirical time complexity of algorithms, that is, the empirical scaling of running time with input size, using a statistically rigorous approach based on resampling statistics. Then, I discuss how automated algorithm configuration can be used to optimise empirical scaling behaviour, using a combination of generic algorithm configuration procedures, racing and statistical testing techniques.

References

1. Holger H. Hoos, Thomas Stützle. *On the empirical scaling of run-time for finding optimal solutions to the travelling salesman problem.* European Journal of Operational Research 238:87–94.
2. James Styles, Holger H. Hoos. *Using Racing to Automatically Configure Algorithms for Scaling Performance.* Proceedings of GECCO 2013, pp. 551–558.

3.14 Modelling and Optimization of Empirical Algorithm Performance

Frank Hutter (Universität Freiburg, DE)

Joint work of Hutter, Frank; Hoos, Holger; Leyton-Brown, Kevin; Xu, Lin
Main reference F. Hutter, L. Xu, H. H. Hoos, K. Leyton-Brown, "Algorithm runtime prediction: Methods & evaluation," Artificial Intelligence, 206(Jan. 2014):79–111, 2014.
URL http://dx.doi.org/10.1016/j.artint.2013.10.003

Algorithm developers and end users often face questions like the following:

- Which parameter setting should I use to optimize my algorithm's empirical performance?
- How does performance depend on the setting of a certain parameter?
- What characteristics distinguish easy from hard problem instances?
- Which of two (or more) available algorithms will perform best on a given new instance?

We describe fully formalized domain-independent methods to answer these questions based on machine learning and optimization techniques. We illustrate the power of these automated methods on a variety of domains, ranging from combinatorial problems (SAT and mixed integer programming) to machine learning (where our methods enable an automatic optimization over the combined space of machine learning algorithms or deep learning architectures and their hyperparameters).

References

1 Frank Hutter, Lin Xu, Holger Hoos, Kevin Leyton-Brown (2014). *Algorithm runtime prediction: Methods and evaluation.* Artificial Intelligence Journal (AIJ) 206:79–111.
2 Frank Hutter, Holger Hoos, Kevin Leyton-Brown (2011). *Sequential Model-Based Optimization for General Algorithm Configuration.* Proceedings of the 5th Learning and Intelligent Optimization Conference (LION 2011).
3 Chris Thornton, Frank Hutter, Holger Hoos, Kevin Leyton-Brown (2013). *Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms.* Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2013).

3.15 Relational Linear Programming

Kristian Kersting (TU Dortmund, DE)

Joint work of Kersting, Kristian; Mladenov, Martin; Tokmakov, Pavel
Main reference K. Kersting, M. Mladenov, P. Tokmakov, "Relational Linear Programs," arXiv:1410.3125v1 [cs.AI, 2014.
URL http://arxiv.org/abs/1410.3125v1

We propose relational linear programming, a simple framework for combing linear programs (LPs) and logic programs. A relational linear program (RLP) is a declarative LP template defining the objective and the constraints through the logical concepts of objects, relations, and quantified variables. This allows one to express the LP objective and constraints relationally for a varying number of individuals and relations among them without enumerating them. Together with a logical knowledge base, effectively a logical program consisting of logical facts and rules, it induces a ground LP. This ground LP is solved using lifted linear programming. That is, symmetries within the ground LP are employed to reduce its dimensionality, if possible, and the reduced program is solved using any off-the-shelf LP solver. In contrast to mainstream LP template languages like AMPL, which features a mixture of declarative and imperative programming styles, RLP's relational nature allows a more intuitive representation of optimization problems over relational domains. We illustrate this empirically by experiments on approximate inference in Markov logic networks using LP relaxations, on solving Markov decision processes, and on collective inference using LP support vector machines.

3.16 The Algorithm Selection Problem: Standard Data Format and Tools

Lars Kotthoff (University College Cork, IE)

The Algorithm Selection Problem is attracting increasing attention from researchers and practitioners from a variety of different backgrounds. After decades of fruitful applications in a number of domains, a lot of data has been generated and many approaches tried, but the community lacks a standard format or repository for this data. Furthermore, there are no standard implementation tools. This situation makes it hard to effectively share and compare different approaches and results on different data. It also unnecessarily increases the initial threshold for researchers new to this area.

In this talk, I will present Aslib, a data format specification and benchmark repository for algorithm selection problems. I will then introduce and briefly demonstrate LLAMA, a modular and extensible toolkit implemented as an R package that facilitates the exploration of a range of different portfolio techniques on any problem domain. LLAMA is integrated with Aslib and is able to work with any algorithm selection scenario in the specified data format.

3.17 Interactive redescription mining using Siren – a demonstration

Pauli Miettinen (MPI für Informatik – Saarbrücken, DE)

Joint work of Galbrun, Esther; Miettinen, Pauli
Main reference E. Galbrun, P. Miettinen, "Interactive Redescription Mining," in Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD'14), pp. 1079–1082, ACM, 2014.
URL http://dx.doi.org/10.1145/2588555.2594520

Exploratory data analysis consists of multiple iterated steps: a data mining method is run on the data, the results are interpreted, new insights are formed, and the resulting knowledge is utilized when executing the method in a next round, and so on until satisfactory results are obtained.

We focus on redescription mining, a powerful data analysis method that aims at finding alternative descriptions of the same entities. We demonstrate Siren, a tool for interactive redescription mining. It is designed to facilitate the exploratory analysis of data by providing a seamless environment for mining, visualizing and editing redescriptions in an interactive fashion, supporting the analysis process in all its stages.

Simultaneously, Siren exemplifies the power of the various visualizations and means of interaction integrated into it, techniques that reach beyond the task of redescription mining considered here, to other analysis methods.

Figure 1 shows an example screenshot from Siren. In the left and back, we see a list of redescriptions explaining areas of Europe on one hand by the mammal species that occupy them, and on the other hand, by the areas' bioclimatic conditions (temperature and rainfall). On the right and front, we see one redescription visualised in a map, where purple means areas where both descriptions hold, red means areas where moose lives but the bioclimatic conditions do not hold, and blue means areas where the bioclimatic conditions hold but moose does not live.

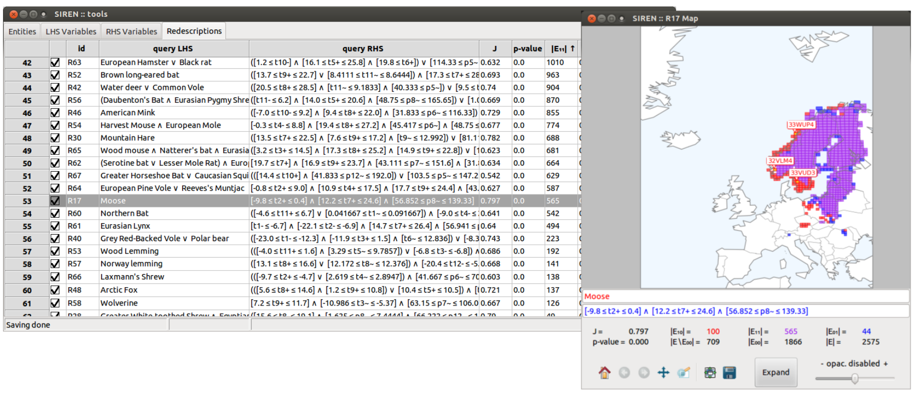

Figure 1 Screenshot of the Siren program.

3.18 Structured Mining Tasks with CP-based frameworks

Benjamin Negrevergne (KU Leuven, BE)

Joint work of Negrevergne, Benjamin; Dries, Anton; Guns, Tias; Nijssen, Siegfried
Main reference B. Negrevergne, A. Dries, T. Guns, S. Nijssen, "Dominance programming for itemset mining," in Proc. of the 2013 IEEE 13th Int'l Conf. on In Data Mining (ICDM'13), pp. 557–566, IEEE, 2013; pre-print available from author's webpage.
URL http://dx.doi.org/10.1109/ICDM.2013.92
URL http://people.cs.kuleuven.be/~benjamin.negrevergne/documents/2013_icdm_dominance_programming.pdf

A long term goal of the pattern mining community has been to design a unified framework for the variety of pattern mining tasks that are available. Recently De Raedt et al. have proposed to use constraint programming (CP) as a unifying framework and have shown that it can be used to elegantly formalize and solve many itemset mining tasks (pattern mining, where patterns are sets). Whether the same technique can be used to formalize more complex mining tasks such as sequence or graph mining is a topic of interest.

In this talk we present our experience with constraint programming for sequence mining. We show that although basic sequence mining tasks can be formalized in constraint programming, the resulting models suffer from several limitations that are not easy to solve in the CP framework. To address these problems, we introduce a more general framework called relational constraint programming which combines ideas from constraint programming and ideas from the relational algebra. Relational constraint programming is a language that resemble the relational algebra but uses constraint programming techniques for evaluating the queries. Finally we show many sequence mining problems can be elegantly formalized and solved in relational constraint programming.

3.19 Relational Constraint Programming

Siegfried Nijssen (Universiteit leiden, NL; KU Leuven, BE)

Joint work of Negrevergne, Benjamin; Dries, Anton; Guns, Tias; Nijssen, Siegfried
Main reference B. Negrevergne, A. Dries, T. Guns, S. Nijssen, "Dominance programming for itemset mining," in Proc. of the 2013 IEEE 13th Int'l Conf. on In Data Mining (ICDM'13), pp. 557–566, IEEE, 2013; pre-print available from author's webpage.
URL http://dx.doi.org/10.1109/ICDM.2013.92
URL http://people.cs.kuleuven.be/~benjamin.negrevergne/documents/2013_icdm_dominance_programming.pdf

This talk argues that a significant number of data mining and machine learning problems can not be formalized as pure constraint satisfaction or optimization problems. Many problems also require a definition of *preference* between solutions, require some form of *aggregation*, and require that some variables ignored. To formalize and solve such problems well-known constraint programming systems, such as Numberjack and the MiniZinc toolchain, do not provide the required support. We propose a new formalism, relational constraint programming, together with solution strategies for this formalism, that do allow for formalizing and solving this wide range of data mining and machine learning problems.

3.20 Structured Learning Modulo Theories

Andrea Passerini (University of Trento – DISI, IT)

Joint work of Passerini, Andrea; Paolo, Campigotto; Roberto, Battiti; Stefano, Teso; Roberto, Sebastiani
Main reference S. Teso, R. Sebastiani, A. Passerini, "Structured Learning Modulo Theories," arXiv:1405.1675v2 [cs.AI], 2014.
URL http://arxiv.org/abs/1405.1675v2

Modelling problems containing a mixture of Boolean and numerical variables is a long-standing interest of Artificial Intelligence. However, performing inference and learning in hybrid domains is a particularly daunting task. The ability to model this kind of domains is crucial in 'learning to design' tasks, that is, learning applications where the goal is to learn from examples how to perform automatic de novo design of novel objects. In this talk I will present Structured Learning Modulo Theories, a max-margin approach for learning in hybrid domains based on Satisfiability Modulo Theories, which allows to combine Boolean reasoning and optimization over continuous linear arithmetical constraints. I will present a number of artificial and real world scenarios showing the potential of the approach.

3.21 Probabilistic Conditional Preferences

Francesca Rossi (University of Padova, IT)

Joint work of Cornelio, Cristina; Goldsmith, Judy; Grandi, Umberto; Mattei, Nicholas; Rossi, Francesca; Venable, Brent

PCP-nets generalise CP-nets to model conditional preferences with probabilistic uncertainty. In this paper we use PCP-net to compactly model a collection of CP-nets, thus using probabilistic uncertainty to reconcile (possible conflicting) preferences expressed by a group of agents. We then study two main tasks: finding an optimal outcome which best represents the preferences of the group of agents, and answering dominance queries. Theoretical and experimental analysis allows us to find efficient and accurate algorithms to perform both tasks.

3.22 Building bridges between data mining and constraint programming

Lakhdar Sais (Artois University – Lens, FR)

Joint work of Jabbour, Said; Sais, Lakhdar; Salhi, Yakoub; Takeaki, Uno
URL http://www.cril.univ-artois.fr/decMining/

This talk, we overview our contribution to data mining and more generally to the cross-fertilization between data mining, constraint programming and propositional satisfiability (http://www.cril.univ-artois.fr/decMining/). We will focus on two contributions. First, we show how propositional satisfiability can be used to model and solve problems in data mining. As an illustration, we present a SAT-based declarative approach for enumerating top-k (closed, frequent) itemsets in transactional databases. Secondly, we discuss the potential contribution of data mining to propositional satisfiability. In this context, we present a first application of data mining to compress Boolean formulas conjunctive normal form.

This work is supported by the ANR DAG project "Declarative approaches for enumerating interesting patterns" (http://liris.cnrs.fr/dag/).

References

1 Said Jabbour and Lakhdar Sais and Yakoub Salhi. *The Top-k Frequent Closed Itemset Mining Using Top-k SAT Problem*. In proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML/PKDD'13, pp. 403–418, Prague, LNAI 8188, Springer, September 23–27, 2013.

2 Said Jabbour and Lakhdar Sais and Yakoub Salhi and Takeaki Uno. *Mining-based Compression Approach of Propositional Formulae*. In Twenty-Second ACM International Conference on Information an Knowledge Management CIKM'2013 , pp. 289–298, October 27th to November 1st, San Francisco, 2013.

3.23 C10: Timed, Probabilistic (concurrent) constraint programming

Vijay A. Saraswat (IBM TJ Watson Research Center – Hawthorne, US)

Over the last twenty years, the theory of concurrent constraint programming has seen significant progress. CCP offers a declarative framework (in the "computation as deduction" paradigm) for concurrent, constraint based computation that subsumes backward-chaining (pure Prolog) and forward- chaining computations, timed computations, and probabilistic computations. We are now developing C10, a concrete language in this framework intended to be compiled into (concurrent, distributed) X10, and support combinatorial problem solving and (probabilistic) analytic computations on big data.

3.24 Learning Constraint Value

Michele Sebag (University of Paris South XI, FR)

Active Learning (AL) and Experiment Design (ED) are problems where one is looking for constraints: these constraints serve to prune the space of candidate hypotheses.

The constraint is made of a question, and the oracle's answer: the oracle gives the label of the instance (AL), or runs the experiment and reports the behavior of the system under identification (ED).

When the oracle is expensive, you are given a budget k: find the most informative k instances /experiments.

This talk discussed:

- how to formalize active learning and experiment design on a budget as reinforcement learning problems (learn to sequentially sample the instance or the experiment space)
- how to find an approximation of the optimal sampling strategy, using Monte-Carlo Tree Search.

3.25 Constraints To Specify Data

Arno Siebes (Utrecht University, NL)

Patterns one observes in a data set can be seen as constraints on possible further data sets one samples from the same distribution. For example, if the vast majority of your customers buy both diapers and beer today, you expect them to do the same tomorrow. Different from normal constraint problems, however, not all solutions – i. e., data sets that satisfy a given set of constraints – are equally good. For example, if our constraints tell us that the data should be smaller than 1 and bigger than -1, a data set that only has values in $[0, 1]$ is not a very satisfying solution; it is not typical for the set of all solutions.

The obvious solution to this problem is to define a probability distribution on the set of all possible data sets such that the more typical data sets have a high probability. Since the collection of all possible data sets is rather large, we do not define the probability explicitly, but implicitly through a generative model, i. e., a model that can generate each possible data set.

The models we consider consist of a small number of of pairs (m_i, p_i) in which m_i is a pattern and p_i is (proportional to) the probability that m_i is used in the generation of a tuple in a new data set. Stipulating that the elements of a model are independent than leads to the simple generative procedure of picking elements proportional to the p_i.

The quality of such a generative model M lies in how likely it is to generate a data set D' that resembles the original data set D. That is, if we use M to generate, say, 1000 data sets and we "hide" D in that collection: how easy is it to identify D. If we can find D is (considerably) less than a logarithmic number of steps, M is a bad model otherwise it is a good model. This informal idea can be formalised through the Minimum Description Length principle.

As usual in MDL, our goal is to find the smallest good model, but algorithms to achieve this goal are not part of this talk.

3.26 OpenML: Networked science in machine learning

Joaquin Vanschoren (TU Eindhoven, NL)

Joint work of Vanschoren, Joaquin; van Rijn, Jan N.; Bischl, Bernd; Torgo, Luis
Main reference J. Vanschoren, J. N. van Rijn, B. Bischl, L. Torgo, "OpenML: networked science in machine learning," SIGKDD Explorations, 15(2):49–60, 2013.
URL http://www.kdd.org/sites/default/files/issues/15-2-2013-12.pdf
URL http://www.openml.org

Today, the ubiquity of the internet is allowing new, more scalable forms of scientific collaboration. Networked science uses online tools to share and organize data on a global scale so that scientists are able to build directly on each other's data and techniques, reuse them in unforeseen ways, and mine all data to search for patterns.

OpenML.org is a place to share and reuse machine learning data sets, tools and experiments. It offers web services for easy integration in many machine learning systems and is readily integrated in R, Weka, Rapidminer and others. It helps researchers win time by reusing and automating machine learning experiments, and gain more credit for their work by making it more easily discoverable and easily reusable.

Moreover, OpenML helps scientists and students to explore different machine learning techniques, find out which are most useful in their work, and collaborate with others to analyze scientific data online. OpenML can itself also learn from the experiments of many researchers, and use algorithm selection and optimization techniques to help people make better use of machine learning in their work.

3.27 Correlation Constraints

Toby Walsh (NICTA – Sydney, AU)

Joint work of Walsh, Toby; Hebrard, Emmanuel; Kizilan, Zeynep

Correlation constraints ensure that solutions are correlated or uncorrelated. I motivate the introduction of correlation constraints, and discuss how they can be propagated. Applications including security games where we wish schedules to be uncorrelated, and logistics where we may wish tomorrow's route to be correlated with today's. Whilst computing the correlation between two solutions is easy, propagation of correlation constraints can be more intractable to compute as we are essentially computing all solutions.

4 Working Groups

4.1 Declarative Languages for Machine Learning and Data Mining

Ian Davidson (UC Davis, USA), Luc De Raedt (KU Leuven, BE), Siegfried Nijssen (KU Leuven, BE; and Universiteit Leiden, NL)

Within the seminar, there was a strong interest in declarative languages for machine learning and data mining. Existing machine learning and data mining algorithms are almost always implemented in procedural languages and a natural evolution is for their implementation in a declarative language. A discussion group was formed that was focused on the following dimensions:

- *General vs. specific languages.* It was agreed upon that generic languages were the focus of the discussion group; within such languages, the focus was on identifying the primitives necessary for allowing mining and learning.
- *Modeling vs. programming.* The focus of the discussion group was on declarative modeling languages that allow to specify a mining or learning task, such as minimum square error clustering; programming languages, which allow to specify execution steps, such as those of the $k-$means algorithm, were not considered.
- *Languages vs. environment.* The focus of the discussion group was on languages. Although it was agreed that a programming environment that provides feedback to the programmer, such as a debugger, is useful, it was felt that the scope would be too broad when including environments in the discussion.

The group first identified why a declarative language for data mining and machine learning is desirable. Several benefits were identified for declarative modeling languages:

- these languages allow for quick modeling of problems;
- these languages allow for quick prototyping of solution methods;
- these languages require less lines of code to specify and solve a task;
- these languages allow for solving tasks for which currently no specialized algorithm exists;
- these languages encourage the reproducibility of results;
- these languages force one to think about and identify general principles;
- programmers can learn from the formal problem specifications and use these specifications when developing specialized solvers;
- the primitives within such languages could categorize the area; one could think of building a catalogue of constraints;
- the language can be used as a specification language: one could check later whether results of specialized algorithms are correct for certain specifications.

A key aspect of the modeling language is the choice of primitives that it supports. Possible primitives were considered to be those for specifying:

- Optimization criteria (loss functions) that capture the underlying learning problem;
- The calculation of Kernels and distances;
- Statistical assumptions, for instance, in probability distributions;
- Constraints and preferences that can encode human guidance and problem constraints;
- The underlying tasks (structure of the problem), including the hypothesis (or function) space and the structure of the output.

A lively discussion was held involving the last aspect.

When restricting the attention to machine learning, two perspectives were considered. One is the one in which machine learning is formalized as finding a function in a hypothesis space such that a loss function is optimized. Each function in the hypothesis space is a function that maps an instance in an input space to an element in an output space, such as for example a class label. A language for machine learning in this case should include primitives for specifying a hypothesis space and a loss function.

The other perspective is that in which machine learning is formalized as a constraint optimization problem, that is, finding an assignment to variables such that constraints and optimization critera are satisfied on these variables. Primitives in such a language should allow to specify variables, their domains, constraints and optimization criteria.

Even though differently focused, it was concluded that these perspectives may be mapped to each other: the variables and domains in one perspective correspond to a hypothesis space in the other perspective, while the loss function corresponds to the optimization criterion. Input and output spaces are encoded across variables, domains, constraints and optimization criteria.

Another discussion revolved around the differences between machine learning and data mining. The argument was made that machine learning and data mining differ from each other as machine learning is specifically focused on finding functions from input to output spaces, while data mining does not always have an output space. This lack of output space reflects the more exploratory nature of data mining.

An inventarisation was made of tasks that could be studied using a declarative data mining and machine learning language. Simple tasks that have already been studied include $k-$means clustering, spectral clustering, itemset mining and decision tree induction.

More difficult tasks include: manifold learning, learning SVMs, structured output prediction, graph clustering or frequent graph mining and label propagation.

As a challenge it was identified to model a number of these problems in existing declarative modeling languages. The following representative problems were picked:

- frequent graph mining;
- learning to rank;
- structured output prediction;
- label propagation.

The data mining and machine learning members of the discussion group agreed to provide written specifications of the problems, while the constraint programming members agreed to formalize these problems in existing modeling languages. This challenge will continue after the Dagstuhl seminar. It is expected that these tasks will pose various problems for existing modeling systems. The discussion group is expected to continue over e-mail after the seminar.

4.2 Learning and Optimization with the Human in the Loop

Andrea Passerini (University of Trento – DISI, IT), Michele Sebag (Université Paris Sud, FR)

4.2.1 Preamble

This discussion group actually merged two topics: Learning and Optimization (L & O, Andrea Passerini) and Human in the Loop (H-I-L, Michele Sebag).

4.2.2 Formal background

Let $\mathcal{X}$ define a solution space. Let f, the objective function to be optimized, be defined on $\mathcal{X}$. f, referred to as scoring function or quality function, is usually defined from $\mathcal{X}$ onto $\mathbb{R}$; occasionally it can be defined onto $\mathbb{R}^d$ (multi-objective optimization: quality, robustness, cost).

In L & O, f is mostly learned from data (with possible feedback from the human in the loop); in H-I-L, f is mostly learned from the the human in the loop.

In all applications, the goal is twofold: learn f; find $\arg\max f$. Depending on the applications, the stress is put on the learning or on the optimization task.

Function f is assumed to abstract all there is to know about the decision problem problem details: decision variables, domains, constraints, objectives preferences.

4.2.2.1 Position of the problem

ILO tackles a sequential decision problem: in each iteration
- The ILO agent must build a hypothesis from the available evidence;
- Based on this evidence, it must act (reinforcement learning) and/or ask for additional evidence (active learning) and/or achieve model-based optimization (active optimization).

4.2.2.2 Theory

- Is there a functional gradient that the ILO trajectory follows (as in boosting)?
- Currently, the criteria leading the ILO trajectory are based on: Expected Global Improvement and Continuous Bayesian Optimization [17, 4, 12, 6]; Expected Posterior Utility [16].

4.2.3 Possible applications

- Automated algorithm selection and configuration [9, 8]
- Structured recommendation, optimal design, computational creativity [15]
- Agents/robots (policy learning)

4.2.4 Categories of problems

4.2.4.1 Nature of the ILO input

Feedback to the agent can be provided by a user, some experimental procedure, or the environment itself. Depending on the source, different types of feedback can be conceived:

- quality of candidate solution
- preferences between candidate solutions [16, 11, 5, 1, 18, 10]
- explanations on why a certain solution is not satisfactory
- sketches of desired solution
- feedback on partial solutions, or repairs a solution proposed by the system [11, 10]
- comments on parts of the model [2]
- partial reward

4.2.4.2 Representation of the solution space

Different degrees of complexity can be identified for the solution space, depending on the type of application which can be conceived:

homogeneous vector in the simplest case the solution space is $\mathcal{X} = \mathbb{R}^D$: the goal is to find some optimal vector x in $\mathbb{R}^D$. This is the typical setting in interactive optimization.

heterogeneous vector a more complex setting is when the solution space is restricted according to variable dependencies (e. g. some dependent variables are relevant only for certain values of the variables they depend on). This can be thought of a solution space $\mathcal{X} = \mathbb{R}^D \times \{0, 1\}^{D'}$ where variable dependencies impose restrictions on the relevant portions of the space. An application example is algorithm configuration [9, 7].

structure many optimization tasks require to return a structured object, with possibly both continuous and discrete variables, and relations between them. Application scenarios are in recommendation, optimal design (e. g. in chemistry, molecule maximizing a given property), marketing (bundle of related items − sketch of a house). Here the space of functions f can be formalized in terms of stochastic graph grammars (less expensive, a stochastic context free grammar) or constrained optimization problems.

function in reinforcement learning, the solution space is a functional space. The sought solution is a policy, i. e. a function mapping any state onto an action. Note that the state abstracts everything relevant to decision.

4.2.4.3 Representation of the objective function space

The type of representation for the objective function depends on the type of expected interaction with the user:

- In the white box case, the user can inspect f and provide explicit comments/repair/debug.
- In the black box case, the user only provides feedback on the solution proposed by the ILO agent.
- In the gray box case, the user can comment on parts of the solution, or on parts of f.

4.2.4.4 Learning the representations

If a dataset including real solutions is available, it is possible to learn constraints (e. g. learning from positive only). Another possibility, inspired from continuous language model [3], is to use supervised learning and discriminate actual solutions from lesioned ones. Unsupervised learning techniques can be also employed to learn useful internal representations for the solutions (representation learning). These approaches aim at learning explicit or implicit constraints on f. However, learning f requires some supervised input (feedback).

4.2.5 At the crossroad of ML and Computer Human Interaction

Some issues are at the cross-road of ML and CHI.

4.2.5.1 User profiling

The resolution of the ILO problem will firstly depend on who is the human in the loop and what is her profile: digitally proficient or naive, wants a solution or wants an explanation or both; wants no burden, etc.

4.2.5.2 Human bias

The user's feedback is bound to be noisy (even more so if preference drift is taking place). There is a systematic bias (manifest e. g. in Robotics application; e. g. Thomaz et al. mention that some users say "it's good" even when undeserved, to "encourage" the agent :-))

4.2.5.3 Non stationarity

Related to the two above is the fact that the user might change her mind along the process (preference drift).

4.2.5.4 The Crowd in the Loop: CIL

There might be more than one user. Having a crowd in the loop increases the noise.

4.2.5.5 ILO on a budget

Querying the user is analogous to optimizing an expensive function [14]. You might want to limit a priori the number of queries, the number of experiments [13], etc. Related questions:
- What is the stopping criterion (when the user says so?)
- How far are we from the goal?

4.2.6 Interface and friendliness

The ILO agent should not jump here and there in the solution space: the queries must display some stability; the user must be and feel in control. Still, experiments in robotics suggest that the process is more efficient if the learning partner is leading the interaction.

4.2.7 Autonomy and Safety

Two settings must be distinguished.
1. The safe case is when the ILO agent can act and ask the user's (or the world's) feedback.
2. The critical case is when the ILO agent must ask for feedback first and act if allowed.

Many cases are in the intermediate region: the ILO agent should ask for feedback in the early interaction phase; but at some points it should know when it can act autonomously.

References

1 Riad Akrour, Marc Schoenauer, Michèle Sebag, and Jean-Christophe Souplet. Programming by Feedback. In *International Conference on Machine Learning*, Pékin, China, June 2014.

2 Ch. Bessiere, R. Coletta, E. Hebrard, G. Katsirelos, N. Lazaar, N. Narodytska, C.-G. Quimper, and T. Walsh. Constraint acquisition via partial queries. In . Rossi, editor, *Proc. IJCAI*. IJCAI/AAAI, 2013.

3 Antoine Bordes, Léon Bottou, Ronan Collobert, Dan Roth, Jason Weston, and Luke Zettlemoyer. Introduction to the special issue on learning semantics. *Machine Learning*, 94(2):127–131, 2014.

4 Eric Brochu, Nando de Freitas, and Abhijeet Ghosh. Active preference learning with discrete choice data. In *Advances in Neural Information Processing Systems*, 2007.

5 J. Fürnkranz, E. Hüllermeier, W. Cheng, and S.-H. Park. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine Learning*, 89(1–2):123–156, 2012.

6 Roman Garnett, Yamuna Krishnamurthy, Xuehan Xiong, Richard Mann, and Jeff G. Schneider. Bayesian optimal active search and surveying. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 1239–1246, New York, NY, USA, 2012. ACM.

7 Holger H. Hoos. Programming by optimization. *Commun. ACM*, 55(2):70–80, February 2012.

8 F. Hutter, H. Hoos, and K. Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *Proc. of ICML-14*, 2014. To appear.

9 Frank Hutter, Lin Xu, Holger H. Hoos, and Kevin Leyton-Brown. Algorithm runtime prediction: Methods & evaluation. *Artificial Intelligence*, 206(0):79–111, 2014.

10 A. Jain, B. Wojcik, T. Joachims, and A. Saxena. Learning trajectory preferences for manipulators via iterative improvement. In *Neural Information Processing Systems (NIPS)*, pp. 575–583, 2013.

11 W. Bradley Knox, Peter Stone, and Cynthia Breazeal. Training a robot via human feedback: A case study. In *Social Robotics*, October 2013.

12 Daniel James Lizotte. *Practical Bayesian Optimization.* PhD thesis, Edmonton, Alta., Canada, 2008. AAINR46365.

13 A. Llamosi, A. Mezine, F. D'Alché-Buc, V. Letort, and M. Sebag. Experimental design in dynamical system identification: A bandit-based active learning approach. In *ECML PKDD*, volume 8725-2, pp. 306–321, 2014.

14 I. Loshchilov, M. Schoenauer, M. Sebag, and N. Hansen. Maximum Likelihood Online Adaptation of Hyper-parameters in CMA-ES. In Th. Bartz-Beielstein et al., editor, *PPSN XIII*, pp. 70–79. LNCS 8672, Springer Verlag, 2014. Best Paper Award.

15 Andrea Passerini Thomas Gärtner, Roman Garnett, editor. *Constructive Machine Learning*, 2013.

16 Paolo Viappiani and Craig Boutilier. Optimal bayesian recommendation sets and myopically optimal choice query sets. In *Advances in Neural Information Processing Systems 23 (NIPS)*. MIT press, 2010.

17 J. Villemonteix, E. Vázquez, M. Sidorkiewicz, and E. Walter. Global optimization of expensive-to-evaluate functions: an empirical comparison of two sampling criteria. *J. Global Optimization*, 43(2-3):373–389, 2009.

18 Aaron Wilson, Alan Fern, and Prasad Tadepalli. A bayesian approach for policy learning from trajectory preference queries. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pp. 1133–1141. Curran Associates, Inc., 2012.

4.3 Meta-Algorithmic Techniques

Holger Hoos (University of British Columbia, CA)

Meta-algorithmic techniques are methods that operate on algorithms with the goal to achieve performance improvements on specific classes of inputs (or problem instances). Examples for such techniques are algorithm configuration, algorithm selection, algorithm scheduling, parallel algorithm portfolios and restart strategies. For all of these, there are per-set (or per-distribution) and per-instance variants, as well as offline, online and dynamic/adaptive variants. Closely related to such meta-algorithmic design techniques are performance prediction methods, which enable many of them, particularly selection.

Apart from clarifying concepts and terminology, the session focussed on compiling an overview of existing techniques, some discussion of what users actually want, challenges and open questions, and directions for future work. In the following, I give a brief overview of each of these topics.

Existing techniques and systems

General per-set algorithm configuration systems

- ParamILS (based on stochastic local search), see http://www.cs.ubc.ca/labs/beta/Projects/ParamILS/
- I/F-Race (based on iterative racing), see http://iridia.ulb.ac.be/irace/
- GGA (based on a genetic algorithm), see https://wiwi.uni-paderborn.de/dep3/entscheidungsunterstuetzungssysteme-und-operations-research-jun-prof-dr-tierney/research/source-code/#c37765
- SMAC (based on sequential model-based optimisation), see http://www.cs.ubc.ca/labs/beta/Projects/SMAC/

Of these, SMAC is the only general system that is based on machine learning, in the sense that it builds a model that can predict performance on previously unseen parameter settings. There are similar approaches that are more limited to a machine learning context, e. g. https://github.com/berndbischl/mlrMBO, aspects of which might perhaps be integrated into future general configuration procedures. AClib (http://www.aclib.net is a useful library of algorithm configuration scenarios, which provides a single interface to ParamILS, I/F-Race and SMAC.

Per-instance configuration systems

- ISAC (IBM) – not publically available due to licensing issues; earlier version available as Matlab and R implementation, see https://sites.google.com/site/yurimalitsky/downloads.
- Hydra (UBC) – not currently publicly available (but public release is planned for the near future)

Algorithm selection systems

- SATzilla (UBC), see http://www.cs.ubc.ca/labs/beta/Projects/SATzilla/; currently only publicly available in a Matlab implementation that is hard to use; a more usable Java implementation should be available soon.
- CSHC (IBM) – not publically available due to licensing issues;
- LLAMA, see https://bitbucket.org/lkotthoff/llama
- Claspfolio, see http://www.cs.uni-potsdam.de/claspfolio/
- SNNAP, see https://sites.google.com/site/yurimalitsky/downloads
- ARS, see http://hal.archives-ouvertes.fr/docs/00/92/28/40/PDF/techrep_ARS.pdf; system should be available soon, could be extended to instance-specific configuration.

ASlib (http://www.aslib.net) is a library of algorithm selection scenario and simple baseline algorithms.

Parallel portfolio systems

- parHydra – currently not publically available (but this may change in the future)

Online learning systems

- Continuous Search in Constraint Programming (Arbelaez, Hamadi, Sebag, ICTAI 2010); also Alejandro Arbelaez' PhD thesis
- Jan N. van Rijn, Geoffrey Holmes, Bernhard Pfahringer, Joaquin Vanschoren:Algorithm Selection on Data Streams. Discovery Science 2014:325–336.

Lifelong learning systems

- REACT – currently not publically available
- elSAC – currently not publically available

Note: Offline algorithm configuration can be used for online configuration (lifelong learning) by collecting instances over time and using an offline configurator on the resulting sets, perhaps using some expiry or down-weighting mechanism for older instances.

Parameter importance techniques that can be used to assess which of a given set of parameters of an algorithm impact performance most, individually or in combination with others

- Functional ANOVA, see http://www.automl.org/fanova – designed to be easy to use.
- Ablation analysis, see http://www.cs.ubc.ca/labs/beta/Projects/Ablation/
- Forward selection, see http://www.cs.ubc.ca/labs/beta/Projects/EPMs/ – currently only Matlab code, quite difficult to use.

What users want

Two example scenarios were discussed that illustrate some aspects of what users of meta-algorithmic techniques might be particularly interested in:

- Given a SAT instance, run the best SAT parameterized algorithm for this. Currently only available for algorithm selection (e. g., SATzilla as it ran in the SAT competition).
- Lifelong learning: improve performance over time (see links to systems above).

Research questions and challenges

The following research questions and challenges were identified:
- How to produce insight into given algorithms and problems, using meta-algorithmic techniques?
- Better features – there is motivation for domain experts to develop these features, e. g., impact on performance, insights into what works where and why.
- Probing features for big data: based on subsets of data, short runs on data, ...
- Given a problem X, encoded into Y and solved as Y, should we use problem features from X, from Y, or both? (See: Proteus: A Hierarchical Portfolio of Solvers and Transformations, Hurley, Kotthoff, Malitsky, O'Sullivan, CPAIOR 2014)
- Training data is very important; we need better principles for building / selecting effective training sets for configuration, selection, etc.
- How to measure how good a training set is (as a training set)? Once we can measure this, we might be able to optimise it.

Automated modelling

Many formalisms (MIP, CSP, SAT, SMT, ASP, ...) permit various representations or encodings of a given problem. Automating this modelling tasks is an important topic, and the group spent some time discussing existing work and open questions in this areas.

- Dominion (St. Andrews) is a generic system that supports automated modelling; it is quite powerful, but at this point quite difficult to use. It currently does not use much machine learning.
- When is reformulation or transformation needed? How should it best be done? (Additional constraints? Transfer to different modelling paradigm, e. g., MIP to CSP?)
- To which extent can existing configuration / selection techniques be used to help with modelling?
- To which extent is modelling akin to search in a combinatorial space?
- Important: symmetry breaking at the model and instance level
- For MIP: a central goal in reformulations is to obtain tighter linear relaxations (should draw on existing work on reformulations, for example Sherali-Adams reformulations, other work on extended formulations; also Dantzig-Wolfe, see http://www.or.rwth-aachen.de/gcg/
- Spiral: synthesis for Fourier transforms (its space contains the approaches from hundreds of journal papers), see http://www.spiral.net/software/sfft.html
- Franz Brglez (http://www.csc.ncsu.edu/people/brglez) demonstrated that in SAT, renaming variables / reordering clauses can cause major performance loss in high-performance DPLL algorithms.
- Alternatives / variants of models should be evaluated on multiple solvers / solvers configured for those (to avoid bias by solver / solver configuration)
- Potentially very interesting: model selection.

Overall, participants in the session felt that meta-algorithmic techniques were already very useful, but poised to have even larger impact; there was also a general sense that much interesting work remained to be done in this area.

4.4 Big Data

Siegfried Nijssen (KU Leuven, BE; and Universiteit Leiden, NL), Barry O'Sullivan (University College Cork, IE)

This working group considered the relationships between big data, data mining and constraint programming.

4.4.1 What is Big Data?

The first question studied was "what is big data?". IBM have proposed four useful dimensions on which to characterise big data: variety (heterogeneity), volume (scale of data), veracity (uncertainty of data), and velocity (streaming data).[3] Therefore, big data presents a major challenge to data scientists. Methods must capable of working at scale both from a volume and speed perspective and capable of reasoning under uncertainty. Heterogeneity presents significant integration issues. Big data also presents major human-computer interaction challenges to facilitate understanding; explanation and visualisation technologies are, therefore, a major opportunity.

4.4.2 Applications

The list of big data applications discussed included:

- Systems biology: e. g., understanding the differences between organisms.
- Pedigree Learning: understanding relationships between organisms.
- Hyper-spectral imaging: huge amounts of event data.
- Radio-astronomy applications: for instance, 800ns is the minimum frequency for studying gravitational waves, and produces massive amounts of data.
- Twitter: geo-positional data, data about spread of viruses, or learning about preferences or reputation.
- Crowd-sourcing: campaigns where non-experts are asked to enrich data from experts, e. g. biodiversity campaigns.
- Peace-keeping and humanitarian applications: data that is collected as a result of the Nuclear Test-ban Treaty, to measure isotopes in the air, ultra-sound, seismographs, etc.; what makes this data interesting is also that different sensors are measuring at different speeds.
- Carpooling, GPS, etc.
- Machine translation.
- Physics: data such as from the ice-cube project, measuring neutrinos, and data from CERN, most of which will be analyzed after the accelerators have stopped.
- Data from the sky-survey project.
- Data collected for predictive maintenance in factories, bridges, etc.

[3] http://www.ibmbigdatahub.com/tag/587

4.4.3 Privacy and Ethical Issues

There was considerable discussion around privacy and ethical issues. It was observed that anonymising data sets is very difficult due to the availability of data that can be joined and cross-referenced. The example was given of a join of the Netflix and IMDB databases. A particularly interesting example related to the "How Unique are You?" project at Harvard.[4] That project showed how almost 90% of the population of one particular US state could be uniquely identified by combining two publicly available data-sets.

An important issue is that of adversary models: what do you want to keep private, and what is the adversary who wishes to violate one's privacy assumed to be capable off?

4.4.4 Techniques

A discussion focused on techniques that can be applied on big data. One question that was raised was which complexity bounds should hold on big data. Promising methods that were mentioned are based on *core-point analysis*. Within a dataset these methods determine core points, and only these core points are fed into a standard algorithm.

An alternative is to use a sampling approach.

The main application in which both techniques do not work is that of outlier detection due to computational complexity issues.

4.4.5 Data & Optimization

Several topics were discussed at the intersection of big data and optimization:
- Finding a good visualization is an optimization problem.
- Finding a good summarization of big data is an optimization problem.
- Increasingly, also optimization models can be big, leading to 'big models', which pose a challenge for current solvers.
- The creation of CSPs may involve combining a large number of different data sources (e. g., from the Internet), and hence the use of big data. See for instance Open Constraint Satisfaction by Faltings et al. In such big data contexts, solution techniques become more 'repair'-based, e. g. local search, large neighbourhood search; these techniques can also operate in parallel.

4.5 Towards Killer Applications

Siegfried Nijssen (KU Leuven, BE; and Universiteit Leiden, NL), Barry O'Sullivan (University College Cork, IE)

4.5.1 What is a Killer Application?

The first discussion item was: what is a killer application? Two definitions may be used:
- Something that is used by a large population of people, but is not necessarily a sophisticated tool addressing a complex problem.
- An application that has a large impact on a well-defined area.

[4] http://aboutmyinfo.org/index.html

4.5.2 Examples of Killer Applications

These were discussed as examples of well-established killer applications of current machine learning, data mining and optimization technology:

- Predicting disease transmission from geo-tagged micro-blog data, as performed by Henry Kautz, Adam Sadilek, Henry A. Kautz, Vincent Silenzio (AAAI, 2012).
- Sudoku, since it has had a large impact on creating an awareness of reasoning about constraints amongst the general population.
- Optical character recognition.
- Spam filtering.
- Google machine translation.
- Recommender systems on websites.
- Massive open online courses.
- Automatic tagging of photos (for machine learning), and associated with it: retrieve by content (what it was) rather than address (where you put it) and personal information management.
- User modelling.

4.5.3 Competitions

Competitions can be useful to encourage progress in an area. Examples of such competitions that were deemed to have had major impact included:

- competitions run by ChaLearn, http://www.chalearn.org/.
- competitions on Kaggle, http://www.kaggle.com.
- the configurable SAT Solver Challenge, http://aclib.net/cssc2014/.
- the AutoML Challenge (run by ChaLearn), http://codalabtest.cloudapp.net/competitions/762.

4.5.4 What Applications would have a Massive Effect (in 5 years)?

Applications with such an impact might be:

- Applications for protecting/managing personal privacy.
- Evidence-based medicine.
- Platforms for facilitating the expression of preferences, preference aggregation, learning over time, collaboration.
- Robot scientists.
- Sports, including video game tournaments.
- Synthetic literature and music generation.
- Advertising.
- Smart homes.
- Alhzeimers treatment planning.
- Companions for the elderly – robotic companions.
- Complex services – air traffic.
- Autonomous cars and traffic management.
- Computational Sustainability.

Participants

- Hendrik Blockeel
KU Leuven, BE
- Jean-François Boulicaut
INSA – Lyon, FR
- Ken Brown
University College Cork, IE
- Bruno Crémilleux
Caen University, FR
- James Cussens
University of York, GB
- Krzysztof Czarnecki
University of Waterloo, CA
- Thi-Bich-Hanh Dao
University of Orleans, FR
- Ian Davidson
Univ. of California – Davis, US
- Luc De Raedt
KU Leuven, BE
- Marc Denecker
KU Leuven, BE
- Yves Deville
University of Louvain, BE
- Alan Frisch
University of York, GB
- Randy Goebel
University of Alberta, CA
- Valerio Grossi
University of Pisa, IT
- Tias Guns
KU Leuven, BE
- Holger H. Hoos
University of British Columbia – Vancouver, CA
- Frank Hutter
Universität Freiburg, DE
- Kristian Kersting
TU Dortmund, DE
- Lars Kotthoff
University College Cork, IE
- Pauli Miettinen
MPI für Informatik – Saarbrücken, DE
- Mirco Nanni
ISTI-CNR – Pisa, IT
- Benjamin Negrevergne
KU Leuven, BE
- Siegfried Nijssen
KU Leuven, BE
- Barry O'Sullivan
University College Cork, IE
- Andrea Passerini
University of Trento – DISI, IT
- Francesca Rossi
University of Padova, IT
- Lakhdar Sais
Artois University – Lens, FR
- Vijay A. Saraswat
IBM TJ Watson Research Center – Hawthorne, US
- Michele Sebag
University of Paris South XI, FR
- Arno Siebes
Utrecht University, NL
- Guido Tack
Monash Univ. Melbourne, AU
- Yuzuru Tanaka
Hokkaido University, JP
- Joaquin Vanschoren
TU Eindhoven, NL
- Christel Vrain
University of Orleans, FR
- Toby Walsh
NICTA – Sydney, AU

Report from Dagstuhl Seminar 14412

Globalizing Domain-Specific Languages

Edited by

Betty H. C. Cheng[1], Benoit Combemale[2], Robert B. France[3], Jean-Marc Jézéquel[4], and Bernhard Rumpe[5]

1 **Michigan State University – East Lansing, US,** chengb@cse.msu.edu
2 **University of Rennes, FR,** benoit.combemale@irisa.fr
3 **Colorado State University – Fort Collins, US,** france@cs.colostate.edu
4 **University of Rennes, FR,** Jean-Marc.Jezequel@irisa.fr
5 **RWTH Aachen, DE,** Rumpe@se-rwth.de

Abstract

This report documents the program and the outcomes of the Dagstuhl Seminar 14412 "Globalizing Domain-Specific Languages" held in October 2014.

Complex, data-intensive, cyber-physical, cloud-based etc. systems need effective modeling techniques, preferably based on DSLs to describe aspects and views. Models written in heterogeneous languages however need to be semantically compatible and their supporting individual tools need to be interoperable. This workshop discusses possible and necessary forms of interoperation their benefits and drawbacks and in particular whether there is a general pattern on coordination, composition and interoperation possible. Main goal was to establish a research programme towards such techniques.

Seminar October 5–10, 2014 – http://www.dagstuhl.de/14412
1998 ACM Subject Classification I.6.5 [Simulation and Modeling]: Model Development – Modeling methodologies, D.3.0 [Programming Languages]: General – Standards, D.2.13 [Software Engineering]: Reusable Software – Domain engineering, Reuse models
Keywords and phrases Modelling, Domain Specific Language, Software, Coordination, Globalization, Heterogeneous Complex Systems, DSL, UML, Composition
Digital Object Identifier 10.4230/DagRep.4.10.32
Edited in cooperation with Katrin Hölldobler

1 Executive Summary

Betty H. C. Cheng
Benoit Combemale
Robert B. France
Jean-Marc Jézéquel
Bernhard Rumpe

Model Driven Engineering (MDE) aims to reduce the accidental complexity associated with developing complex software-intensive systems, through the development of technologies that enable developers to systematically create, evolve, analyse, and transform various forms of abstract system models.

Current MDE language workbenches, in both academia and industry, support the development of Domain-Specific Modeling Languages (DSMLs) that can be used to create models that play pivotal roles in different development phases. Language workbenches such as EMF, Metaedit+ or MPS support the specification of the abstract syntax, concrete syntax and the static and dynamic semantics of a DSML. These workbenches aim to address the needs of DSML developers in a variety of application domains.

The development of modern complex software-intensive systems often involves the use of multiple DSMLs that capture different system aspects. In addition, models of the system aspects are seldom manipulated independently of each other. System engineers are thus faced with the difficult task of relating information presented in different models. Current DSML development workbenches provide good support for developing independent DSMLs, but provide little or no support for integrated use of multiple DSMLs. The lack of support for explicitly relating concepts expressed in different DSMLs (incl., syntax and semantics) makes it very difficult for developers to reason about information spread across models describing different system aspects.

Supporting coordinated use of DSMLs leads to what we call the globalization of modeling languages, that is, the use of multiple modeling languages to support coordinated development of diverse aspects of a system.

Discussions during the seminar will focus on how multiple heterogeneous modeling languages (or DSMLs) will need to be related to determine how different aspects of a system influence each other. We have identified three forms of relationships among DSMLs that can be used as a starting point for discussions: interoperability, collaboration, and composition. These forms of language integration will need to address challenging issues that arise from the heterogeneity of modeling languages. Relationships among the languages will need to be explicitly defined in a form that corresponding tools can use to realize the desired interactions. Requirements for tool manipulation is thus another topic that will be discussed in the seminar.

The goal of the seminar was to develop a research program that broadens the current DSML research focus beyond the development of independent DSMLs to one that provides support for globalized DSMLs. In the globalized DSMLs vision, integrated DSMLs provide the means for teams working on systems that span many specialized domains and concerns to determine how their work on a particular aspect influences work on other aspects.

Working Groups

In the seminar we started the following four working groups which are producing results during the workshop and compiling them into a State-Of-The-Art report afterwards:

- **Group 1a** Motivating Use Cases for the Globalization of DSLs Definition of the main scenarios motivating the globalization of DSLs
- **Group 1b** Conceptual Model of the Globalization of Domain-Specific Languages Definition of the common vocabulary and foundations of the globalization of DSLs
- **Group 2** Globalized Domain Specific Language Engineering Challenges of the globalization of DSLs from the language designer point of view
- **Group 3** Domain Globalization: Using Languages to Support Technical and Social Coordination Challenges of the globalization of DSLs from the language user point of view

2 Table of Contents

3 Overview of Talks

3.1 Globalizing DSLs through Contextualized Modeling

Colin Atkinson (Universität Mannheim, DE)

As software systems increase in size and complexity, and are expected to cope with ever more quantities of information from ever more sources, there is an urgent and growing need for a more view-oriented approach to software engineering. Views allow stakeholders to see exactly the right information, at exactly the right time, in a way that best matches their capabilities and goals. Domain-specific languages are a key foundation for supporting views by allowing them to display their contents in a customized way, but the current generation of software language engineering technologies do not go far enough. In particular, they currently lack the ability to convey the precise relationship between the information shown in a view and the information it is a view of. They also focus on describing how model elements should be visualized but provide little or no support for describing how stakeholders should edit and interact with them.

The premise of this talk is that software language engineering technologies need to evolve to support an enhanced approach to modeling in which model content can be set in context relative to the underlying source from which it is derived – an approach we refer to as "contextualized modeling". These technologies would then be more accurately characterized as "view engineering" technologies rather than "language engineering" technologies since they would support all aspects of view definition, including the context in which the content is to be interpreted and the mechanisms by which model elements are to be visualized and edited. Some of the key additional capabilities that the current generation of language engineering technologies need to support in order to become globalized, viewpoint engineering languages include:

Enriched Designation. The most important context information in a view is its model elements' location in the three key hierarchies of the underlying information model – the classification hierarchy, the inheritance hierarchy and the containment (i. e. ownership) hierarchy. These are supported to various degrees in today's language engineering technologies through a mix of explicit symbolism and location-defining designators (a.k.a. headers) in model elements. However, they are not supported in a uniform and consistent way, and are often severely limited in what they can express. In particular most contemporary language engineering technologies only allow one level of classification to be expressed at a time. Fully contextualized modeling requires a comprehensive, systematic and deep designation notation which allows a model element's exact location in each hierarchy to be expressed in its designator.

Explicit Elision Symbolism. Since views almost always convey only a subset of the information contained in the underlying model, an important requirement in viewpoint engineering is to support the description of what things are not included in a view, as well as the description of what things are. This is a challenging task since it involves subtle interactions between explicit omission statements (e. g. "…" in UML generalization sets), explicit completeness statements (e. g. complete and disjoint in UML generalization sets) and background "world" assumptions (e. g. "open world" versus "closed world" assumption). Fully contextualized modeling therefore requires comprehensive and systematic support for elision, both in the form of explicit elision symbols and elided model element designators.

Explicit Derivation Symbolism. As well as omitting information from the underlying subject of a view it is possible to derive new information that the subject does not explicitly contain. Such derivation operations can be driven by the application of basic characterization relationships such as inheritance and classification (e. g. subsumption) or by more complex inference operations based on the principles of logic. In both cases, contextualized modeling must incorporate the ability to express what information in a view has been derived and what information has been explicitly asserted by a human modeller. This is important for resolving conflicts and signalling the weight that should be given to the information represented within views.

Language Symbiosis. Domain-specific representations of information have the advantage that they are optimized for particular classes of stakeholders or communities of experts, whereas general-purpose languages have the advantage that they are widely known and can represent information in quasi-standard ways. In order to enjoy both benefits simultaneously, contextualized models should be represented by highly flexible, symbiotic languages that allow different visualizations of model elements to be mixed and interchanged at will.

Viewpoint Environment Definition. A user's experience of a view is determined not only by the way in which its contents are displayed, but also by the way in which the user can interact with the model and, when it is editable, input information. This impacts all aspects of the environment in which the view is displayed, including the menu items, the pallets of predefined types and models elements and the range of operations that can be applied to the content (e. g. checking, printing, persisting etc.). The engineering of viewpoints therefore involves much more than just the engineering of languages it also involves the definition of the associated interaction experience.

In this talk the vision of contextualized modeling is introduced and the key ingredients needed to turn the current generation of software language engineering technologies into fully globalized viewpoint-engineering technologies explained.

3.2 Globalizing DSLs: overall consistency and large scale collaboration

Cedric Brun (Obeo – Nantes, FR)

Domain specific languages have shown their efficiency in designing more precisely and easing the creation of tools. The system definition ends up being split by concern which mostly maps to the type of stakeholders involved, but in the end we need to have a view of the consistency of the system globally.

DSLs have relationships in between them, a DSL might expose some aspects through interfaces formalizing how others can use or extend it. This lack of formalization leads to shortcomings in regard to the technical integrations of those DSLs and this gets even more complex when collaboration gets involved.

As a provider of technologies enabling the use of DSLs at a large scale tackling these challenges has a huge potential, complexity could still be managed thanks to using languages which are domain specifics yet more stakeholders could be involved and the whole process would require less coordination.

3.3 Globalization of Modeling Languages: A Formal Semantics Approach

Barrett Bryant (University of North Texas, US)

The goals of this research are to 1) formalize the semantics of Domain-Specific Modeling Languages (DSMLs), 2) use semantic formalization for automatic generation of model-driven engineering software tools, and 3) use semantic formalization and tool generation to facilitate DSML composition. This will allow automating many tasks that are currently done ad hoc in a manual hand-crafted manner.

3.4 Hybrid systems modeling and simulation: challenges and solutions

Benoit Caillaud (INRIA Bretagne Atlantique – Rennes, FR)

Hybrid systems combine continuous and discrete time dynamics, expressed in a single language, or as combination of several of several dedicated languages. Two key challenges regarding hybrid systems are:

- Integration of discrete and continuous time models at a semantics level. How can one co-simulate a system model combining models with radically different semantics: discrete time dynamical systems on one hand, and continuous time dynamical systems on the other hand. The discrete time dynamics is often expressed in a data flow or automata based language, while the continuous dynamics results from a system of ordinary or algebraic differential equations (resp. ODEs and DAEs). Several techniques can be used to address this problem, depending on the overall system architecture and the assumptions that can be made on the overall system behaviour. These techniques range from simple asymmetric co-simulation methods, where time is handled by the numeric differential equation solver, to more involved techniques, for example, waveform relaxation.
- A second challenge is compositionality and modular compilation of acausal continuous time models. They are often expressed using algebraic differential equations (DAEs), where the data flow orientation of an incomplete model may depend on its environment. Hence, generating simulation code for a component, without knowing its precise environment, is a difficult problem. The problem becomes even more severe when considering hybrid systems with DAEs, found for example in Modelica models. The main reason is that, not only the dataflow orientation may change dynamically, but the differentiation index may change, depending on the discrete state of a model.

3.5 Application-Driven Globalization of Modeling Languages

Betty H. C. Cheng (Michigan State University – East Lansing, US)

Our society is now demanding software for engineered systems at levels of complexity that transcend historical precedents and the state of the art in software development concepts, methods and tools. Major challenges arise due to many factors: unprecedented functions and complexity; significant uncertainty about requirements [1], designs and components; the inability due to scale to impose centralized control on system definition, development, deployment, operation and evolution; hard performance requirements; and continual scattered failure due to accidents and attacks. Such systems have recently been called Ultra-Large-Scale (ULS) Systems. New knowledge and methods are needed in both traditional and emerging areas of software research to enable successful development of ULS systems. Neither industry nor the academic community is well-positioned to develop this knowledge by itself.

Software and its corresponding models are key enablers of ULS systems—the source of new behaviors as well as the flexibility to adapt under uncertainty and change —- but also, due to shortcomings in our knowledge, the key impediment to developing such systems. Difficulties are being experienced today in a wide range of application areas, including defense, transportation systems, financial systems, medical-based systems, etc. Demands for function, safety, flexibility, responsiveness, availability, security, privacy and integration far outstrip current knowledge and engineering capabilities.

A specific subset of ULS systems are high-assurance systems (HAS), those systems designed to tolerate component failures, and even direct attacks, in order to continue operation and preserve system integrity. The study of high-assurance systems is inherently multidisciplinary, requiring collaboration of educators and researchers from a wide variety of scientific disciplines and application domains [5]. Moreover, academic researchers need to work closely with industrial partners to ensure that solutions address the scale and complexity found in the real world, and to facilitate the transition of research results into practice.

Designing and implementing high-assurance systems typically requires the integration of several enabling computing technologies (e. g., model-driven software engineering [2, 9], sensor networks, autonomic computing [7]) within a particular application domain (e. g., transportation systems, telecommunication networks, electronic medical records, digital supply chain). In addition, human factors play a critical role, since most realistic applications involve a blend of humans and machines. Each of these orthogonal dimensions may involve a wide range of modeling languages, all of which need to be integrated in order to provide the overall functionality with the required assurance [4, 6]. Model integration challenges exist at the structural, semantic, and application domain levels [3]. Beyond the technical challenges, it is important for the research to address the usability of the model-based systems for a wide range of stakeholders, from domain experts to policy analysts and other types of decision-makers [8].

References

1 Jon Whittle, Pete Sawyer, Nelly Bencomo, Betty H. C. Cheng and Jean-Michel Bruel, RELAX: a language to address uncertainty in self-adaptive systems requirement, In Requirements Engineering, Vol. 15, Issue 2, pages 177–196, Springer London, 2010

2 Heather Goldsby and Betty H. C. Cheng, Automatically Discovering Properties That Specify the Latent Behavior of UML Models, MoDELS (1), pages 316–330, 2010

3 B. H. C. Cheng, A. Ramirez, and P. K. McKinley, Harnessing evolutionary computation to enable dynamically adaptive systems to manage uncertainty, Combining Modelling and Search-Based Software Engineering (CMSBSE), Keynote for 2013 1st International Workshop on, 2013

4 Andres J. Ramirez, Betty H. C. Cheng, Nelly Bencomo, and Pete Sawyer, Relaxing Claims: Coping with Uncertainty While Evaluating Assumptions at Run Time., MoDELS, 2012

5 Robert B. France, James M. Bieman, Sai Pradeep Mandalaparty, Betty H. C. Cheng, and Adam C. Jensen, Repository for Model Driven Development (ReMoDD)., ICSE,2012

6 Andres J. Ramirez, Erik M. Fredericks, Adam C. Jensen and Betty H. C. Cheng, Automatically RELAXing a Goal Model to Cope with Uncertainty., SSBSE,2012
7 Philip K. McKinley, Betty H. C. Cheng, Andres J. Ramirez, and Adam C. Jensen, Applying evolutionary computation to mitigate uncertainty in dynamically-adaptive, high-assurance middleware., In J. Internet Services and Applications, pages 51–58, 2012
8 Gunter Mussbacher, Daniel Amyot, Ruth Breu, Jean-Michel Bruel, Betty H. C. Cheng, Philippe Collet, Benoit Combemale, Robert B. France, Rogardt Heldal, James H. Hill, Jörg Kienzle, Matthias Schöttle, Friedrich Steimann, Dave R. Stikkolorum, and Jon Whittle, The Relevance of Model-Driven Engineering Thirty Years from Now, Model-Driven Engineering Languages and Systems – 17th International Conference, MODELS 2014, Valencia, Spain, September 28 – October 3, 2014. Proceedings pages 183–200, 2014
9 Adam C. Jensen, Betty H. C. Cheng, Heather Goldsby, and Edward C. Nelson, A Toolchain for the Detection of Structural and Behavioral Latent System Properties, Model Driven Engineering Languages and Systems, 14th International Conference, MODELS 2011, Wellington, New Zealand, October 16–21, 2011. Proceedings, pages 683–698,2011
10 Peter H. Feiler, Kevin Sullivan, Kurt C. Wallnau, Richard P. Gabriel, John B. Goodenough, Richard C. Linger, Thomas A. Longstaff, Rick Kazman, Mark H. Klein, Linda M. Northrop, and Douglas Schmidt, Ultra-Large-Scale Systems: The Software Challenge of the Future, Software Engineering Institute, Pittsburgh, 2006

3.6 Experience integrating DSLs and Formal Methods for Coordinating Vehicles

Siobhan Clarke (Trinity College Dublin, IE)

Our overall research goal was to define a Domain Specific Language for applications that would take advantage of a middleware that caters for vehicles coordinating in real-time. We wanted to allow an application developer build an application (e. g., an intersection collision avoidance system or a managed highway) using our vehicle coordination protocol where we could then verify the safety of the application. It was intended that the safety checking would be achieved by integrating a formal method language. However, in our experience, this was challenging primarily because our requirements were not matched by any potentially related formal language. This is a general challenge for globalising languages, as their integration may still not address all requirements in the "new", global context.

3.7 Globalization of Domain Specific Modelling Languages

Tony Clark (Middlesex University, GB)

The globalisation of domain specific modelling languages can be achieved through an understanding of language composition. This involves the specification, implementation and deployment of multiple languages and their associated artefacts including syntax, semantics, documentation, methods etc. In order to understand and achieve globalisation it is fruitful to take a component-based view of languages where a language component defines a collection of interfaces that expose those parts of a language specification or implementation that are necessary to integrate it with other languages. At the time of writing it is not clear how to express or use such language components and this is an important research are that needs to be addressed. A useful approach may be to develop a common meta-language that can be used to articulate languages as components. Furthermore, a common framework based on such meta-concepts may be useful for globalisation where existing languages and their models can be wrapped in order to conform to the requirements of the framework.

3.8 Language Engineering Workbench

Benoit Combemale (University of Rennes, FR)

In the software and systems modeling community, research on domain-specific modeling languages (DSMLs) is focused on providing technologies for developing languages and tools that allow domain experts to develop system solutions efficiently. Unfortunately, the current lack of support for explicitly relating concepts expressed in different DSMLs makes it very difficult for software and system engineers to reason about information spread across models describing different system aspects. Supporting coordinated use of DSMLs leads to what we call the globalization of modeling languages. I present a research initiative that broadens the DSML research focus beyond the development of independent DSMLs to one that supports globalized DSMLs, that is, DSMLs that facilitate coordination of work across different domains of expertise.

3.9 Golden Models in Engineering and Formal Methods

Julien DeAntoni (INRIA Sophia Antipolis – Méditerranée, FR)

In many disciplines and for several years, models have been used to abstract the system under study. Depending on the model and its purpose, it brings very different properties ranging from re-usability to analyzability. Consequently, there is usually no single golden model of a system, but there are good models of a system where different dedicated models abstract the system for different purpose. Real time embedded systems are interesting candidates for modeling for two main reasons. On the one hand many properties like timing performance, time-to-market or safety are early and mandatory requirements to be satisfied at all steps of the design. On the other hand, the deployment of such system can target various heterogeneous platforms and this deployment strongly impacts the previously stated requirements.

For some years in the AOSTE team, I am studying how engineering models (e. g., based on UML or Ecore) and formal models (e. g., automaton or marked graph) can take benefits one from the other in order to improve the modeling of real time embedded systems. More precisely my current research focuses on two related topics. First, to enable reasoning on engineering models I put efforts to provide formal models that describe the behavioral semantics of a language. This means providing an adequate meta-language to specify an explicit entity that represents the behavioral semantics and can thus be manipulated. Second, because different models are used to specify correctly a single system, it is important to understand the interaction among these different models and more precisely to understand how the explicit semantics can be used to provide a behavioral interface of the language amenable to reasoning, composition and generative techniques.

3.10 Globalizing Domain-Specific Languages: A few thoughts on the open-world

Thomas Degueule (University of Rennes, FR)

In the past few years, the development of model-driven engineering and associated tools has strengthened the trend supporting domain-specific language development by enabling domain experts to design their own DSL without requiring strong skills in languages or compilers construction. This has led to a widespread use of DSL in many areas, including software and system engineering.

While the facilities available for engineering isolated DSLs are getting closer to maturity, there is little support for comprehension on the interactions between several languages that are used within a single system or software. In an open world, where DSLs are designed independently by (possibly small) groups of developers, the need for "globalization" is increasingly felt.

To support the globalization of DSLs, we advocate the design of precise language interfaces. Language interfaces allow to abstract some of the complexity carried in the implementation of languages, by exposing meaningful information concerning a given aspect of a language and for a specific purpose (e. g. composition or coordination) in an appropriate formalism. We strongly believe that such interfaces will ease the cognitive and technical effort necessary for lifting the limitations of current approaches.

3.11 Social Translucence

Robert B. France (Colorado State University, US)

We are interested in designing systems that support communication and collaboration among large groups of people over computer networks. We begin by asking what properties of the physical world support graceful human-human communication in face-to-face situations, and argue that it is possible to design digital systems that support coherent behavior by making participants and their activities visible to one another. We call such systems "socially translucent systems" and suggest that they have three characteristics-visibility, awareness, and accountability-which enable people to draw upon their social experience and expertise to structure their interactions with one another. To motivate and focus our ideas we develop a vision of knowledge communities, conversationally based systems that support the creation, management and reuse of knowledge in a social context. We describe our experience in designing and deploying one layer of functionality for knowledge communities, embodied in a working system called "Babble," and discuss research issues raised by a socially translucent approach to design.

3.12 The Need for Multilevel DSMLs

Ulrich Frank (Universität Duisburg-Essen, DE)

The construction of DSMLs is facing an essential challenge: On the one hand, it makes sense to develop very specific languages that serve the particular needs of one organization only, because then we can expect it to effectively promote model quality and modelling productivity. On the other hand, global DSMLs promote economies of scale by enabling a wider range of reuse and integration. A multilevel hierarchy of DSML, which is inspired by the actual use of technical languages, enables to successfully address this challenge. On the top level, which corresponds to the concepts introduced in textbooks, global DSML would capture commonalities of a range of more specific DSML. Depending on the size and diversity of the domains to be covered, the number of required levels may vary. Realizing respective hierarchies of DSMLs convincingly demands for adequate modelling concepts and – more challenging – for multilevel (meta) programming languages. In addition to linguistic and technical considerations, there is need to build effective incentives for contributing to the development of DSML hierarchies.

3.13 Towards Families of DSLs

Jean-Marc Jézéquel (University of Rennes, FR)

The engineering of systems involves many different stakeholders, each with their own domain of expertise. Hence more and more organizations are adopting Domain Specific Languages (DSLs) to allow domain experts to express solutions directly in terms of relevant domain concepts. This new trend raises new challenges about designing DSLs, evolving a set of DSLs and coordinating the use of multiple DSLs. In this talk we explore various dimensions of these challenges, and outline a possible research roadmap for addressing them. We detail one of these challenges, which is the safe reuse of model transformations.

3.14 Globalization of modeling languages: definition and challenges

Gabor Karsai (Vanderbilt University, US)

The problem of globalization of modeling languages can be viewed as a system integration problem for modeling languages and tools. Integration of modeling languages deals with the composition of heterogeneous modeling languages, so that the composite has a clear semantics – i. e. the composed models have a meaning, possibly beyond that of the component models. Integration of modeling tools deals with the orchestration of complex tool use cases, where heterogeneous tools work together to execute a complex tasks, e. g. co-simulation, formal verification across multiple models, etc. There three fundamental challenges associated with these topics: (1) determining the relevant properties of the modeling languages and tools that need to be captured in some formalism (e. g. metamodels), (2) determining the composition operators for metamodels that facilitate integration, and (3) assigning operational semantics to the composition of the modeling languages and tools. The first problem deals with developing the structural semantics of a model (or tool) integration language, the second deals with the specific composition operators available in the language, and the third addresses the operational semantics of the model integration language.

3.15 Globalizing Modeling Languages

Marjan Mernik (University of Maribor, SI)

The globalization of modeling languages is a newly emerged term defining a situation where multiple heterogeneous modeling languages are used for describing different aspects of a complex system. These aspects may or may not (partially) overlap. Moreover, there is an urgent need that these heterogeneous modeling languages interact to each other. Therefore, the syntax and semantics of modeling languages must be precisely defined. One form of interaction known from programming languages is language composability, which is a property of language specifications rather than a property of a language itself. They are several known forms of programming language composition: language extension (which subsumes also language restriction), language unification, self-extension, and extension composition. However this classification needs to be extended and adopted for globalization of modeling languages.

3.16 Smart Emergency Response as an Application Use Case

Pieter J. Mosterman (The MathWorks Inc. – Natick, US)

Main reference P. J. Mosterman, D. E. Sanabria, E. Bilgin, K. Zhang, J. Zander, "A Heterogeneous Fleet of Vehicles for Automated Humanitarian Missions," Computing in Science & Engineering, 16(3):90–95, IEEE, 2014.
URL http://dx.doi.org/10.1109/MCSE.2014.58

This presentation explores the cyber-physical systems paradigm in humanitarian missions, in particular in emergency response to natural disasters such as earthquakes. A smart emergency response system is presented to illustrate opportunities and challenges in cyber-physical system applications.

3.17 Domain-Specific Tooling Infrastructure

Oscar M. Nierstrasz (Universität Bern, CH)

Globalization of modeling languages is about enabling the shared use of such languages for domains of common interest. We are working on better tools to rapidly extract models from source code, techniques to adapt development tools to specific domains, and generic DSLs that can be adapted to a variety of different analysis tools. The key challenges we identify are: (1) better environments and workbenches to support DSL engineering, (2) techniques to rapidly develop or adapt tools to specific domains, and (3) techniques to support the integration and coordination of multiple DSLs into software systems.

3.18 Composition of Languages

Bernhard Rumpe (RWTH Aachen, DE)

DSLs need to be modular, reusable and composable. Otherwise we will get a plethora of incompatible and complex languages that are expensive to develop, maintain and evolve. We therefore take the view of component based design, by adapting the mechanisms of modular components and their composition to languages. This means that a "language component" should have a crisp boundary, that encapsulates internals of the language and makes the language accessible only through its interface.

The notion of "language interface" is not easy to assess. Languages certainly exhibit parts of their abstract syntax, which is the main carrier for a language, for composing sub-languages together into more complex ones. However the concrete syntax needs also to be composed if applicable. Furthermore for a precise understanding of the emerging composed language, the semantics, which we like to be given as denotational semantics with semantics domain and semantics mapping needs to be composed as well. This is tricky, as semantics domains can be completely disjoint, overlapping or even identical, but encoded in different carriers. Last but not least, additional information about the symbols defined in a sub-model and exported to another part of a model, which is defined in another sub-language, must be transferable through the language interface.

While we do have a relatively good understanding of concrete compositions of computer languages, I useful and general theory for language composition, adaptation and thus their reuse is still in infancy. We discuss some results and more problems on the concepts of language engineering, including syntax, semantics and tooling that we made while our development of and with the MontiCore language workbench.

3.19 Globalization of DSLs or All about Boxes and Lines

Martin Schindler (MaibornWolff GmbH – München, DE)

Main reference M. Schindler, "Eine Werkzeuginfrastruktur zur agilen Entwicklung mit der UML/P," Dissertation, RWTH Aachen, Aachener Informatik-Berichte – Software-Engineering, Vol. 11, 370 pages, Shaker Verlag, 2012.
URL http://dx.doi.org/10.2370/9783844008647

Globalization of DSLs means composing languages including all involved artifacts. For this a clear definition of a language component and its interface is needed. A language component has to encapsulate all parts of a language like concrete and abstract syntax, validation (e. g., context conditions), and transformations (e. g., code generators). On the other hand, a language interface only includes the necessary information needed for composing a language with other languages.

In [1] it is shown how the UML/P, which is a subset of the UML including Java as an action language, can be developed in such a component based way. All languages of the UML/P (Classdiagrams, Statecharts, Objectdiagrams, Sequencediagrams, OCL, a language for testcases and Java) were developed completely independently of each other including a definition what is required from or provided for other languages. These required and provided interfaces were finally used to compose the UML/P keeping the language components unchanged. In this way the languages of the UML/P can be easily replaced by or reused with other languages.

The UML/P is just an example of language composition. However, at the end the composition of different DSLs should be just about connecting provided and required interfaces of the involved languages and should be as easy as drawing boxes and lines.

References

1 Eine Werkzeuginfrastruktur zur agilen Entwicklung mit der UML/P (A Tooling Infrastructure for the Agile Development with the UML/P) Martin Schindler Shaker Verlag, ISBN 978-3-8440-0864-7. Aachener Informatik-Berichte, Software Engineering Band 11. 2012. http://dx.doi.org/10.2370/9783844008647

3.20 A benchmark for globalizing domain-specific languages

Friedrich Steimann (Fernuniversität in Hagen, DE)

The globalization of domain-specific languages (DSLs) has been achieved if we are able to refactor across DSLs. This is so because refactoring, i. e., the behaviour-preserving change of a system, requires full anticipation of the effect of every change on the behaviour of the system, and hence the joint semantics of the DSLs that are being used.

While being able to refactor across DSLs may seem like a high hurdle, we have found that, once behaviour-preserving changes are mastered in each participating language, cross-language refactoring is little more than identifying the (model or program) elements through which interaction occurs, and translating the changeable properties of these elements from one language to another (so that changes can be propagated).

It seems that the globalization of DSLs requires the same: identification of the elements through which interaction occurs, and full awareness of what changes of these elements translate to in each language. This can be greatly simplified by using a language interoperability infrastructure on which all participating languages are built; in absence of such an infrastructure, it will be much harder (and likely requires effort quadratic in the number of participating languages).

3.21 "Globalizing DSLs": Defining coordination among modeling languages

Juha-Pekka Tolvanen (MetaCase – Jyväskylä, FI)

Domain-Specific Modeling has become increasingly popular in the past decade. These languages allow raising the level of abstraction away from the solution domain (code) to the problem domain with obvious benefits such as improved development productivity and product quality. While typically the domains- specific modeling languages are built for a narrow area within a company the next obvious step is to "globalize" languages so that coordinated use of domain- specific languages becomes possible. We identify some key challenges for research in three contexts: organization, language and technology. In the organizational context often already a single DSL may change organizational tasks, roles and structures. How multiple coordinated ones while influence to organizations and to development processes. In the language context, the coordination must be specified at the level of language specification but it is not clear how currently used metamodeling languages allow to do that. For example OMG's MOF does not even identify "language" so it is questionable how it can then integrate a number of them? Finally, and within the technology context, it is not realistic to expect that all languages can be coordinated within a single tool so what kind of tool integration approach would work among a set of tools?

3.22 A view On the Globalization of Modeling Languages

Antonio Vallecillo (University of Malaga, ES)

DSLs are proliferating, as system modelers and designers are finding them useful for their purposes, and as tool support starts to be readily available for them (editors, validators, code generators, etc.). The previous "one language fits all" approach (e. g. Java, UML) has given path to "one language for each purpose", and this is where the need to make public, coordinate and combine languages (i. e., globalize them) has risen.

"Globalizing modeling languages" was defined in the original GEMOC paper as "The use of multiple languages to support coordinated development of diverse systems aspects". However, I only agree in part with such definition. First, I see it is more adequate for defining what Multi-Viewpoint Modeling is/should be about: "The combination of multiple languages to support coordinated specification, analysis and development of diverse systems aspects"

Thus, in my view, "Globalizing a Modeling Language" means "Making a Modeling Language amenable for integration into a (standard) Multi-Viewpoint Modeling environment."

In this context, it is important to note that (a) Globalized MLs need to be combinable and integrable (b) Interfaces at different levels should be defined , and (c) Standardization should play a key role here.

In our group we have been working in this area, in the context of the Reference Model for Open Distributed Processing (RM-ODP), an ISO & ITU-T international standard that provides a mature framework for the specification of large complex systems, using viewpoints. RM- ODP defines five viewpoints and their associated Viewpoint Languages (VPL), as well as explicit correspondences between the VPLs. This is an example of the coordination and integration of separate languages, focusing on disparate concerns, and using correspondents to relate them.

In addition, we have been working on the combination of DSMLs, studying the problems, issues and challenges involved in this area.

The three major challenges that we see in the globalization of modeling languages are the following. First, there is the need for defining mechanisms, process and tools for the Combination/Integration/Unification of languages (which needs establishing correspondences between them, at all levels: Abstract Syntax, Concrete Syntax and Semantics); and needs to deal with heterogeneous (and not always combinable) semantics. Second, correspondences between metamodels, and between models, needs to be specified in an efficient, correct, usable and maintainable manner, and using different approaches (depending on the use we want to make of them). The third challenge that we want to highlight is about being able to reason about the information expressed across the different models, so that properties of the overall system (including emergent properties) can be inferred, proved or denied.

Tool support is essential in this context for achieving these goals, given the complexity of the domain and of the systems being specified. Without tools any proposed solution will be useless.

3.23 Globalizing Models with the MPS Language Workbench

Markus Völter (Völter Ingenieurbüro, DE)

Developing software often requires using a number of tools and languages. In embedded software, for example, it is common to use C and its IDE, Matlab/Simulink, a number of custom XML files, a requirements management tool such as DOORS and possibly a UML tool and a variant management tool. The integration of such a zoo of tools is often a major source of (accidental) complexity in development projects. The GEMOC initiative addresses this challenge.

Back in the 'good old days' when everything was text files and command line executables running on the unix shell. This approach had two important properties: the infrastructure was extremely generic (unix shell, pipes, text editors) and the actual contents were easily extensible and composable (new text file formats/languages and new command line tools); a productive environment for a given project or domain could easily be built from the generic infrastructure plus a few custom extensions.

Language Workbenches can be used to create (domain-specific) development environments that results in many of the same advantages that we all valued in the unix shell-world. A language workbench is an extremely generic infrastructure that is easily extensible with new languages. It is easy to create domain-specific development tools that can address different concern of the system with suitable abstractions, but are nonetheless very well integrated in terms of syntax, semantics and tooling.

JetBrains MPS is such a language workbench; over the last few years we have used it to build mbeddr, an open source environment optimized for embedded software development. It consists of a set of 50+ C extensions as well as languages for requirements and documentation. It also directly integrates formal verification tools. Connecting to artifacts outside of mbeddr is possible via external references.

Language Workbenches, MPS and systems like mbeddr can be an important contribution for managing the overall complexity of Globalized DSLs.

3.24 DSL related research challenges

Mark van den Brand (TU Eindhoven, NL)

Model Driven Software Engineering is extremely popular in the High Tech Industry. They are facing two challenges. The first challenge is to increase the quality of the code, an ongoing quest. The second challenge is to get grip on the rapidly increasing amount of software. The high tech companies are exploring all options that model driven software engineering offers. Domain specific languages (DSLs) is one of the explored routes. Some of these companies have a whole range of DSLs. So far, these DSLs are developed in house, using different technologies, although that EMF and related technologies is the most important platform being used. Although that the high tech industry is a software intensive industry they do consider themselves not a software industry. They are DSL users and not DSL developers, or formulated differently they are tool users and not tool developers. Some of the high tech companies outsource the DSL and supporting tooling to suppliers.

The DSLs are used to describe different aspects of the high tech systems that are being produced. This means equivalent concepts used in the different DSLs have to be the same over these DSLs. One way of doing this is have a common semantic framework. In close cooperation with ASML we are developing a semantic framework (semantics based language workbench) based on Event-B. This semantic framework is used to give a formal for one of the DSLs used within ASML. The goal of the project is to use this framework to develop new DSLs in the future, in order to ensure the consistency of semantic concepts. The identification of common semantic concepts and capturing them in a semantics based workbench is the first challenge. Related challenges are the correctness of model transformations and modularity of meta models. Each of these challenges is related to the globalization of domain specific languages.

Participants

- Colin Atkinson
Universität Mannheim, DE
- Cedric Brun
Obeo – Nantes, FR
- Barrett Bryant
University of North Texas, US
- Benoit Caillaud
INRIA Bretagne Atlantique – Rennes, FR
- Betty H. C. Cheng
Michigan State University – East Lansing, US
- Tony Clark
Middlesex University, GB
- Siobhán Clarke
Trinity College – Dublin, IE
- Benoit Combemale
University of Rennes, FR
- Julien Deantoni
INRIA Sophia Antipolis – Méditerranée, FR
- Thomas Degueule
University of Rennes, FR
- Robert B. France
Colorado State University, US
- Ulrich Frank
Universität Duisburg-Essen, DE
- Jean-Marc Jézéquel
University of Rennes, FR
- Gabor Karsai
Vanderbilt University, US
- Ralf Lämmel
Universität Koblenz-Landau, DE
- Marjan Mernik
University of Maribor, SI
- Pieter J. Mosterman
The MathWorks Inc. – Natick, US
- Oscar M. Nierstrasz
Universität Bern, CH
- Bernhard Rumpe
RWTH Aachen, DE
- Martin Schindler
MaibornWolff GmbH – München, DE
- Friedrich Steimann
Fernuniversität in Hagen, DE
- Eugene Syriani
University of Alabama, US
- Janos Sztipanovits
Vanderbilt University, US
- Juha-Pekka Tolvanen
MetaCase – Jyväskylä, FI
- Antonio Vallecillo
University of Malaga, ES
- Mark van den Brand
TU Eindhoven, NL
- Markus Völter
Völter Ingenieurbüro, DE

Report from Dagstuhl Seminar 14421

Optimal Algorithms and Proofs

Edited by

Olaf Beyersdorff[1], Edward A. Hirsch[2], Jan Krajíček[3], and Rahul Santhanam[4]

1 **University of Leeds, GB,** o.beyersdorff@leeds.ac.uk
2 **Steklov Institute of Mathematics, St. Petersburg, RU,** edward.a.hirsch@gmail.com
3 **Charles University in Prague, CZ,** krajicek@karlin.mff.cuni.cz
4 **University of Edinburgh, GB,** rsanthan@inf.ed.ac.uk

Abstract

This report documents the programme and the outcomes of the Dagstuhl Seminar 14421 "Optimal algorithms and proofs". The seminar brought together researchers working in computational and proof complexity, logic, and the theory of approximations. Each of these areas has its own, but connected notion of optimality; and the main aim of the seminar was to bring together researchers from these different areas, for an exchange of ideas, techniques, and open questions, thereby triggering new research collaborations across established research boundaries.

Seminar October 12–17, 2014 – http://www.dagstuhl.de/14421
1998 ACM Subject Classification F.1.3 Complexity Measures and Classes, F.2.2 Nonnumerical Algorithms and Problems: Complexity of proof procedures, G.1.2 Approximation
Keywords and phrases computational complexity, proof complexity, approximation algorithms, optimal algorithms, optimal proof systems, speedup theorems
Digital Object Identifier 10.4230/DagRep.4.10.51
Edited in cooperation with Alexander Smal

1 Executive Summary

Olaf Beyersdorff
Edward A. Hirsch
Jan Krajíček
Rahul Santhanam

General Introduction to the Topic

The notion of optimality plays a major role in theoretical computer science. Given a computational problem, does there exist a "fastest" algorithm for it? Which proof system yields the shortest proofs of propositional tautologies? Is there a single distribution which can be used to inductively infer any computable sequence? Given a class of optimization problems, is there a single algorithm which always gives the best efficient approximation to the solution? Each of these questions is a foundational one in its area – the first in computational complexity, the second in proof complexity, the third in computational learning theory, and the last in the theory of approximation.

Consider, as an example, the Boolean Satisfiability (SAT) search problem, which asks, given a Boolean formula, for a satisfying assignment to the formula. Since SAT is NP-complete, being able to tell whether the fastest algorithm for SAT runs in polynomial time would imply a solution to the notoriously hard NP vs P problem, which is far beyond the state of our current knowledge. However, the possibility remains that we can define an *optimal* algorithm which we can guarantee to be essentially the fastest on every instance, even if we cannot rigorously analyze the algorithm. In a seminal paper, Leonid Levin (1973) proved that every NP search problem, and in particular SAT, has an optimal algorithm. It is still unknown whether every decision problem in NP has an optimal algorithm.

In general, given a class of computational artefacts (algorithms/proof systems/distributions) and performance measures for each artefact in the class, we say that an artefact is optimal if it matches the performance of every other artefact in every case. The main questions about optimality is: for which classes of artefacts and under which assumptions do they exist? In case they do exist, how well do they match the performance of other artefacts in the class? How is the existence of optimal artefacts related to other fundamental theoretical questions, such as complexity lower bounds, efficient learnability or approximability?

There have been a number of important recent results about optimality in various computational settings. Prime examples include optimal proof systems and acceptors under advice or in heuristic settings, surprising relations of optimal proof systems to descriptive complexity and parameterized complexity, hierarchy results in various computational settings, and optimal approximation algorithms for constraint satisfaction problems.

Organisation of the Seminar and Activities

The seminar brought together 41 researchers from different areas of computer science and mathematics such as computational complexity, proof complexity, logic, and approximations with complementary expertise, but common interest in different notions of optimality. The participants consisted of both senior and junior researchers, including a number of postdocs and a few advanced graduate students.

Participants were invited to present their work and to communicate state-of-the-art advances. Twenty-two talks of various lengths were given over the five-day workshop. Survey talks of 60 minutes were scheduled prior to workshop, covering the three main areas of computational complexity, proof complexity, and approximations. Most of the remaining slots were filled as the workshop commenced. In addition, during two spontaneously organised open problem sessions – one at the very start and the second, longer one near the end of the workshop – the participants posed a number of open problems across the different disciplines covered by the seminar. The organisers considered it important to leave ample free time for discussion.

Three tutorial talks were scheduled during the first two days in order to establish a common background for the different communities from computational complexity, proof complexity, logic, and approximation that came together for the workshop. The presenters and topics were:

- David Steurer: Survey on Approximations and Optimality
- Olaf Beyersdorff: Optimal Proof Systems – a Survey
- Rahul Santhanam: Hierarchies and Lower Bounds via Optimality – a Survey

The other 19 talks covered a broad range of topics from logic, computational complexity and proof complexity.

The organisers think 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

3 Overview of Talks

3.1 Optimal Proof Systems – a Survey

Olaf Beyersdorff (University of Leeds, GB)

This talk is a survey on optimal proof system. I will not cover any results in detail, but try to present the general picture of what is known and what to expect. The question whether optimal proof systems exist was first raised by Krajíček and Pudlák [9] and has been open since. In the talk I survey

1. Characterizations for the existence of optimal proof systems [1, 3, 4, 9, 10];
2. Sufficient and necessary conditions for their existence [6, 9];
3. Positive results in different models [2, 5, 11];
4. Connections to first-order logic [7, 8].

A longer exposition of the content of the talk is available as a guest post to Hunter Monroe's blog 'Speedup in Computational Complexity'.

References

1 Olaf Beyersdorff, Johannes Köbler, and Jochen Messner. Nondeterministic functions and the existence of optimal proof systems. *Theoretical Computer Science*, 410(38–40):3839–3855, 2009.
2 Olaf Beyersdorff, Johannes Köbler, and Sebastian Müller. Proof systems that take advice. *Information and Computation*, 209(3):320–332, 2011.
3 Olaf Beyersdorff and Zenon Sadowski. Do there exist complete sets for promise classes? *Mathematical Logic Quarterly*, 57(6):535–550, 2011.
4 Yijia Chen and Jörg Flum. From almost optimal algorithms to logics for complexity classes via listings and a halting problem. *J. ACM*, 59(4):17, 2012.
5 Edward A. Hirsch and Dmitry Itsykson. On optimal heuristic randomized semidecision procedures, with application to proof complexity. In *Proc. STACS'10*, pages 453–464, 2010.
6 Johannes Köbler, Jochen Messner, and Jacobo Torán. Optimal proof systems imply complete sets for promise classes. *Information and Computation*, 184(1):71–92, 2003.
7 Jan Krajíček. *Bounded Arithmetic, Propositional Logic, and Complexity Theory*, volume 60 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, 1995.
8 Jan Krajíček. *Forcing with random variables and proof complexity*, volume 382 of *Lecture Note Series*. London Mathematical Society, 2011.
9 Jan Krajíček and Pavel Pudlák. Propositional proof systems, the consistency of first order theories and the complexity of computations. *The Journal of Symbolic Logic*, 54(3):1063–1079, 1989.
10 Jochen Messner. On optimal algorithms and optimal proof systems. In *Proc. STACS'99*, pages 541–550, 1999.
11 Toniann Pitassi and Rahul Santhanam. Effectively polynomial simulations. In *Proc. 1st Innovations in Computer Science*, 2010.

3.2 Total Space in Resolution

Ilario Bonacina (University of Rome "La Sapienza", IT)

Joint work of Bonacina, Ilario; Galesi, Nicola; Thapen, Neil

Consider a resolution refutation of some unsatisfiable formula F. Such refutation could be presented on a blackboard with limited space. Initially the blackboard is empty and at each step of the presentation we can either: write on the blackboard some clause from F; apply the resolution rule to clauses already on the blackboard and write down the clause we get; erase some clause from the blackboard (in order to save space). The refutation ends when we can write the empty clause on the blackboard. The Total Space of F is the minimal size of a blackboard needed to present a refutation of F, where the size of a blackboard is intended to be the number of literals (counted with repetitions) it can contain.

We will show that some constant width formulas in n variables the blackboard must contain at least cn clauses each of width cn, for some constant $c > 0$. Hence require Total Space $\Omega(n^2)$. This result is optimal (up to a constant factor).

3.3 Are There Hard Examples for Frege Systems? – Nearly Twenty Years Later

Samuel R. Buss (University of California – San Diego, US)

We discuss the lack of combinatorial examples of candidate tautologies for exponentially separating Frege and extended Frege systems. Recently, different groups have given quasipolynomial size Frege proofs for determinental identities, Frankl's theorem, and the Kneser-Lovasz tautologies. This talk presents a new proof of the pigeonhole principle which formalizes the Cook Reckhow proofs as quasipolynomial size Frege proofs.

3.4 Majority is Incompressible by $\mathsf{AC}^0[p]$ Circuits

Igor Carboni Oliveira (Columbia University, New York, US)

Joint work of Carboni Oliveira, Igor; Santhanam, Rahul

Razborov/Smolensky (1987) obtained lower bounds on the size of depth-d Boolean circuits extended with modulo p gates computing the Majority function. This result remains one of the strongest lower bounds for an explicit Boolean function. In this work, we obtain information about the structure of polynomial-size Boolean circuits with modulo p gates computing Majority. For instance, we show that for any d, at least $n/((\log n)^{O(d)})$ wires must enter the d-th layer of the circuit, which is essentially optimal. This result follows from the investigation of a more general framework called interactive compression games (Chattopadhyay and Santhanam, 2012), which combines computational complexity and communication complexity, and has applications in cryptography, parametrized complexity and circuit complexity. In this talk, we will discuss new results in this model, and mention a few open problems.

3.5 A Parameterized Halting Problem

Yijia Chen (Shanghai Jiao Tong University, CN)

Joint work of Chen, Yijia; Flum, Jörg
Main reference Y. Chen, J. Flum, "A parameterized halting problem," in H. L. Bodlaender, R. Downey, F. V. Fomin, D. Marx (eds.), "The Multivariate Algorithmic Revolution and Beyond – Essays Dedicated to Michael R. Fellows on the Occasion of His 60th Birthday," LNCS, Vol. 7370, pp. 364–397, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-30891-8_17

The parameterized problem $p{-}Halt$ takes as input a nondeterministic Turing machine M and a natural number n, the size of M being the parameter. It asks whether every accepting run of M on empty input tape takes more than n steps. This problem is in the class XP_{uni}, the class "uniform XP," if there is an algorithm deciding it, which for fixed machine M runs in time polynomial in n. It turns out that various open problems of different areas of theoretical computer science are related or even equivalent to $p{-}Halt \in XP_{uni}$. Thus this statement forms a bridge which allows to derive equivalences between statements of different areas (proof theory, complexity theory, descriptive complexity, . . .) which at first glance seem to be unrelated. As our presentation shows, various of these equivalences may be obtained by the same method.

3.6 Proof Complexity for Quantified Boolean Formulas

Leroy Chew (University of Leeds, GB)

Joint work of Beyersdorff, Olaf; Chew, Leroy; Janota, Mikoláš
Main reference O. Beyersdorff, L. Chew, M. Janota, "Proof complexity of resolution-based QBF calculi," to appear.

Proof systems for quantified Boolean formulas (QBFs) provide a theoretical underpinning for the performance of important QBF solvers. In particular, the calculi Q-resolution and long- distance Q-resolution serve as underlying formalisms for DPLL solvers for QBFs. More recently, calculi based on universal expansion were introduced in order to enable reasoning about expansion-based QBF solvers. These are ∀Exp+Res [3] and its generalisations IR and IRM [1]. However, the proof complexity of these proof systems is currently not well understood and in particular lower bound techniques are missing.

In this talk we exhibit a new and elegant proof technique for showing lower bounds in QBF proof systems based on strategy extraction [2]. This technique provides a direct transfer of circuit lower bounds to lengths of proofs lower bounds. We use our method to show the hardness of a natural class of parity formulas for Q-resolution. Variants of the formulas are hard for even stronger systems as long-distance and universal Q-resolution. With a completely different lower bound argument we show the hardness of the prominent formulas of Kleine Büning et al. for the strong expansion-based calculus IR, thus also confirming the hardness of the formulas for Q-resolution. Our lower bounds imply new exponential separations between two different types of resolution-based QBF calculi: proof systems for DPLL-based solvers (Q-resolution, long- distance Q-resolution) and proof systems for expansion-based solvers (∀Exp+Res and its generalisations IR and IRM). The relations between proof systems from the two different classes were not known before.

References

1. Olaf Beyersdorff, Leroy Chew, and Mikoláš Janota. On unification of QBF resolution-based calculi. In *MFCS*, pages 81–93, 2014.
2. Olaf Beyersdorff, Leroy Chew, and Mikoláš Janota. Proof complexity of resolution-based QBF calculi. In *STACS*. LIPIcs series, 2015.
3. Mikolás Janota and João Marques-Silva. On propositional QBF expansions and Q-resolution. In *SAT*, pages 67–82, 2013.

3.7 On the Success Probability of Polynomial-Time SAT Solvers

Andrew Drucker (University of Edinburgh, GB)

Main reference A. Drucker, "Nondeterministic Direct Product Reductions and the Success Probability of SAT Solvers," in Proc. of the 2013 IEEE 54th Annual Symp. on Foundations of Computer Science (FOCS'13), pp. 736–745, IEEE, 2013.
URL http://dx.doi.org/10.1109/FOCS.2013.84

In one approach to solving NP-complete problems like SAT, we try to design an efficient randomized algorithm that attempts to guess a solution, and that is guaranteed to have success probability better than truly-random guessing (if a solution exists). Such "intelligent random guessing" is at the core of a number of improved exponential-time algorithms for these problems. This was observed by Paturi and Pudlák [1], who found evidence for the limitations of such algorithms.

We further this project. We show that a standard hardness assumption ($NP \notin coNP/poly$) implies the following: For every polynomial-time randomized algorithm attempting to produce satisfying assignments to Boolean formulas, there are infinitely many satisfiable instances on which the algorithm's success probability is nearly-exponentially small. Our proof involves new ideas for the study of average-case complexity in the circuit model.

References

1. R. Paturi, P. Pudlák. *On the Complexity of Circuit Satisfiability.* Proceedings of the forty-second ACM symposium on Theory of computing, pp. 241–250, 2010.

3.8 The Space Complexity of Cutting Planes Refutations

Nicola Galesi (University of Rome "La Sapienza", IT)

Joint work of Galesi, Nicola; Pavel Pudlák; Neil Thapen

We study the space complexity of the cutting planes proof system, in which the lines in a proof are integral linear inequalities. We measure the space used by a refutation as the number of inequalities that need to be kept on a blackboard while verifying it. We show that any unsatisfiable set of inequalities has a cutting planes refutation in space five. This is in contrast to the weaker resolution proof system, for which the analogous space measure has been well-studied and many optimal lower bounds are known.

Motivated by this result we consider a natural restriction of cutting planes, in which all coefficients have size bounded by a constant. We show that there is a CNF which requires super-constant space to refute in this system. The system nevertheless already has an exponential speed-up over resolution with respect to size, and we additionally show that it is stronger than resolution with respect to space, by constructing constant-space cutting planes proofs of the pigeonhole principle with coefficients bounded by two.

We also consider variable space for cutting planes, where we count the number of instances of variables on the blackboard, and total space, where we count the total number of symbols.

3.9 On the Correlation of Parity and Small-Depth Circuits

Johan Håstad (KTH Royal Institute of Technology, SE)

Main reference J. Håstad, "On the Correlation of Parity and Small-Depth Circuits," SIAM J. Computing, 43(5):1699–1708, 2014.
URL http://dx.doi.org/10.1137/120897432

We prove that the correlation of a depth-d unbounded fan-in circuit of size S with parity of n variables is at most $\exp(-\Omega(n/(\log S)^{d-1}))$.

3.10 On Optimal Heuristic Computations and Heuristic Proofs

Dmitry Itsykson (Steklov Institute of Mathematics, St. Petersburg, RU)

Joint work of Itsykson, Dmitry; Hirsch, Edward; Monakhov, Ivan; Nikolaenko, Valeria; Smal, Alexander; Sokolov, Dmitry

An acceptor for a language L is an algorithm that accepts elements of L and does not stop on other inputs. Messner proved that for all good enough (paddable) languages the existence of an optimal acceptor is equivalent to the existence of a p-optimal proof system. We consider a notion of randomized heuristic acceptors that may accept with noticeable probability a small fraction of inputs according to some distribution concentrated on the complement of the language. We show that for every recursively enumerable langauge L and polynomial-time samplable distribution concentrated on the complement of L there exists an optimal randomized heuristic acceptor. Sometimes it is possible to make a construction deterministic. For example for a language of the images of an injective function $f_n : \{0,1\}^n \to \{0,1\}^{n+1}$ there exists an optimal deterministic heuristic algorithm. Sometimes it is also possible to eliminate errors: there exists an average-case optimal randomized acceptor for graph non-isomorphism.

In the heuristic setting the proof of the equivalence between optimal acceptor and p-optimal proof systems fails. However a heuristic proof system is an interesting concept. We give some examples of short heuristic proofs that have no known short classical counterparts.

3.11 QBF Solving and Proof Systems

Mikoláš Janota (INESC-ID, Lisbon, PT)

Joint work of Janota, Mikoláš; Klieber, William; Marques-Silva, Joao; Clarke, Edmund
Main reference M. Janota, W. Klieber, J. Marques-Silva, E. Clarke, "Solving QBF with Counterexample Guided Refinement," in Proc. of the 15th Int'l Conf. on Theory and Applications of Satisfiability Testing (SAT'12), LNCS, Vol. 7317, pp. 114–128, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-31612-8_10

Deciding Quantified Boolean Formulas (QBFs) is interesting both theoretically and practically. QBFs are amenable to solving and theoretical analysis due to its canonic structure . At the same time, they enable expressing a wide range of problems as the decision problem is PSPACE complete. In this talk we will look at a recent method for solving QBF, which gradually expands the given formula and invokes a SAT solver in a blackbox fashion. This approach has proven to be highly competitive compared to existing ones. We will briefly discuss a proof system that corresponds to this solving algorithm.

3.12 New Lower and Upper Bounds on Circuit Complexity

Alexander S. Kulikov (Steklov Institute of Mathematics, St. Petersburg, RU)

In the first part of the talk, we will show how SAT-solvers can help to prove stronger upper bounds on the Boolean circuit complexity. Roughly, the main idea is that circuits for some functions are naturally built from blocks of constant size. E.g., the well-known circuit that computes the binary representation of the sum of n input bits is built from n full adders and has size $5n$. One can then state the question "whether there exist a block of smaller size computing the same function" in terms of CNF- SAT and then ask SAT-solvers to verify this. Using this simple approach we managed to improve the upper bound for the above mentioned function to $4.5n$. This, in particular, implies that any symmetric function has circuit size at most $4.5n + o(n)$. We will also present improved upper bounds for some other symmetric functions.

In the second part we will present much simpler proofs of currently best known lower bounds on Boolean circuit complexity. These are $3n - o(n)$ for the full binary basis [Blum, 1984] and $5n - o(n)$ for the binary basis where parity and its complement are excluded [Iwama, Morizumi, 2002]. The properties of the functions under consideration allow us to prove the stated lower bounds with almost no case analysis.

3.13 Narrow Proofs May Be Maximally Long

Massimo Lauria (KTH Royal Institute of Technology, SE)

Joint work of Atserias, Albert; Lauria, Massimo; Nordström, Jakob
Main reference A. Atserias, M. Lauria, J. Nordström, "Narrow Proofs May Be Maximally Long," in Proc. of the 2014 IEEE 29th Conf. on Computational Complexity (CCC'14), pp. 286–297, IEEE, 2014.
URL http://dx.doi.org/10.1109/CCC.2014.36

We prove that there are 3-CNF formulas over n variables refutable in resolution in width w that require resolution proofs of size n^w. This shows that the simple counting argument that any formula refutable in width w must have a proof in size n^w is essentially tight. Moreover, our lower bound extends even to polynomial calculus resolution (PCR) and Sherali-Adams, implying that the corresponding size upper bounds in terms of degree and rank are tight as well. In contrast, the formulas have Lasserre proofs of constant rank and size polynomial in both n and w.

3.14 An Observation on Levin's Algorithm and a New (?) Application to Matrix Multiplication

Jochen Messner (Ulm, DE)

We present a simple observation on Levin's algorithm which allows an efficient implementation for example on Turing machines. Then we use Freyvald's randomized matrix multiplication test together with Levin's method to obtain an optimal probabilistic matrix multiplication algorithm.

3.15 Speedup and Noncomputability

Hunter Monroe (IMF, Washington, US)

Speedup broadly is the nonexistence of an optimal algorithm under some partial order. The presentation will consider whether speedup exists for "natural" computational problems such as multiplying integers or matrices and not only for Blum's artificially constructed languages. The goal will be to direct attention toward nonexistence rather than existence of optimal algorithms. The talk will: (1) consider worst case speedup for integer and matrix multiplication; (2) note a connection with monotone-nonmonotone gap for Boolean circuits; (3) examine possible infinitely often speedup for the complement of bounded halting (and for coNP-complete languages and for proof systems) and whether better algorithms be easily produced; and (4) discuss a possible relationship between the properties "has no best algorithm" and "has no algorithm at all".

3.16 On Some Problems in Proof Complexity

Pavel Pudlák (Academy of Sciences, Prague, CZ)

We are interested in open problems about the relation of complexity and provability. For most of these statements, it seems that one answer is more plausible than the other. Therefore we rather talk about conjectures. A prototype of such a conjecture is the one that says that there is no finitely axiomatized consistent theory S such that for every finitely axiomatized consistent theory one can construct proofs of $Con_S(n)$ in polynomial time. Here $Con_S(n)$ denotes the consistency of S restricted to proofs of length at most n. The conjectures $P \neq NP$ and $NP \neq coNP$ can also be viewed as such conjectures, because they can be stated in terms of propositional proof systems.

The conjecture that we studied so far can be classified in two ways: (1) deterministic/nondeterministic, (2) universal/existential. The main universal conjectures are comparable and so are the main existential conjectures. Thus the conjectures form two branches. We introduce two new conjecture. One is the Σ_1^b finite reflection principle, which is a natural strengthening of finite consistency mentioned above. The second one is *Herbrand consistency search.* The reason for introducing Herbrand consistency search is to get a conjecture related to consistency also in the existential branch of conjectures.

The strongest conjecture in the universal branch is the conjecture saying that there is no complete disjoint NP pair. Similarly, the strongest conjecture in the existential branch is the conjecture saying that there is no complete disjoint $coNP$ pair. We have not been able to find a natural conjecture that would imply both conjectures.

3.17 On the AC^0 Complexity of Subgraph Isomorphism

Benjamin Rossman (National Institute of Informatics, Tokyo, JP)

Joint work of Rossman, Benjamin; Li, Yuan; Razborov, Alexander

Let P be a fixed graph (hereafter called a "pattern"), and let Subgraph(P) denote the problem of deciding whether a given graph G contains a subgraph isomorphic to P. We are interested in AC^0-complexity of this problem, determined by the smallest possible exponent $C(P)$ for which Subgraph(P) possesses bounded-depth circuits of size $n^{C(P)+o(1)}$. Motivated by the previous research in the area, we also consider its "colorful" version Subgraph$_{col}(P)$ in which the target graph G is $V(P)$-colored, and the average-case version Subgraph$_{ave}(P)$. Defining $C_{col}(P)$ and $C_{ave}(P)$ analogously to $C(P)$, our main contributions can be summarized as follows.

1. $C_{col}(P)$ coincides with the tree-width of the pattern P within a logarithmic factor. This shows that the previously known upper bound by Alon, Yuster, Zwick is almost tight.
2. We give a characterization of $C_{ave}(P)$ in purely combinatorial terms within a multiplicative factor of 2. This shows that the lower bound technique of Rossman is essentially tight, for any pattern P whatsoever.
3. We prove that if Q is a minor of P then Subgraph$_{col}(Q)$ is reducible to Subgraph$_{col}(P)$ via a linear-size monotone projection. At the same time, we show that there is no monotone

projection whatsoever that reduces Subgraph(M_3) to Subgraph($P_3 + M_2$) (P_3 is a path on 3 vertices, M_k is a matching with k edges, and "+" stands for the disjoint union). This result strongly suggests that the colorful version of the subgraph isomorphism problem is much better structured and well-behaved than the standard (worst-case, uncolored) one.

3.18 Characterizing the Existence of Optimal Proof Systems and Complete Sets for Promise Classes

Zenon Sadowski (University of Białystok, PL)

Joint work of Sadowski, Zenon; Beyersdorff, Olaf
Main reference O. Beyersdorff, Z. Sadowski, "Do there exist complete sets for promise classes?" Mathematical Logic Quarterly, 57(6):535–550, 2011.
URL http://dx.doi.org/10.1002/malq.201010021

We investigate the following two questions:

Q1: Do there exist optimal proof systems for a given language L?
Q2: Do there exist complete problems for a given promise class C?

For concrete languages (such as TAUT or SAT) and concrete promise classes (such as UP, disjoint NP-pairs etc.) these questions have been intensively studied during last years, and a number of characterizations have been obtained. Here we provide new characterizations for Q1 and Q2 that apply to almost all promise classes C and languages L, thus creating a unifying framework for the study of these questions. More specifically, we introduce the notion of a promise complexity class representable in a proof system (captured by a proof system). We express the promise condition of a class in a language L and then use a proof system for L to verify that a given Turing machine satisfies the promise.

3.19 Hierachies and Lower Bounds via Optimality: A Survey

Rahul Santhanam (University of Edinburgh, GB)

I survey work on hierarchy theorems and circuit lower bounds, which uses ideas from optimal algorithms. This work includes hierarchy theorems for probabilistic time with advice due to Barak and Fortnow & myself, and my work on circuit lower bounds for MA with advice.

3.20 Disjoint NP-Pairs and Propositional Proof Systems

Alan Selman (SUNY – Buffalo, US)

Joint work of Glasser, Christian; Hughes, Andrew; Selman, Alan; Wisiol, Nils

This talk surveys results on disjoint NP-pairs, propositional proof systems, function classes, and promise classes – including results that demonstrate close connections that bind these topics together. We illustrate important links between the questions of whether these classes have complete objects and whether optimal proof systems may exist.

3.21 Examples of Heuristic Proofs

Dmitry Sokolov (Steklov Institute of Mathematics, St. Petersburg, RU)

Joint work of Sokolov, Dmitry; Itsykson, Dmitry

In this talk we consider heuristic proof systems and give non-trivial examples of proof systems of this kind. We give an example of a distributional problem (Y, D) that is in the complexity class $HeurNP$ but if NP is not equal to $coNP$ then Y is not in NP, and if $(NP, PSamp)$ is not contained in $HeurBPP$ then (Y, D) is not in $HeurBPP$.

For a language L and a polynomial q we define a language L_q composed of pairs (x, r) where x is an element of L and r is an arbitrary binary string of length $q(|x|)$. If $D = \{D_n\}$ is an ensemble of distributions on strings, let $[D, U]$ be a distribution on pairs (x, r), where x is distributed according to D_n and r is uniformly distributed on strings of length $q(n)$. We show that for every language L in AM there is a polynomial q such that for every distribution D concentrated on the complement of L the distributional problem $(L_q, [D, U]_q)$ has a polynomially bounded heuristic proof system. Since graph non-isomorphism (GNI) is in AM, the above result is applicable to GNI.

4 Open Problems

The seminar hosted two open problem sessions: the first immediately after the introduction on Monday morning, thus giving participants the opportunity to state problems they would like to discuss with others during the week, and the second one towards the end of the workshop on Thursday evening, reflecting on material presented during the week. The problems presented in these two sessions include:

1. **Andrew Drucker**
   - Let $PC(\varphi)$ denotes a proof length of φ in some propositional proof system Π. Is there a sequence of tautologies $\varphi_1(x_1, \ldots, x_n), \ldots, \varphi_t(x_1, \ldots, x_n)$, s.t. $PC(\varphi_1, \ldots, \varphi_t) = \omega(\max_i PC(\varphi_i))$?
2. **Nicola Galesi**
   - Can CP^* (cutting-plane proof system with polynomially bounded coefficients) refute every unsatisfiable CNF using constant space?
   - Is it possible to refute every unsatisfiable CNF in CP with linear total space?
   - Devise better lower bounds for CP^2 (cutting-plane proofs with coefficients bounded by 2).

   Background information on these problems can be found in [5].
3. **Andreas Goerdt**
   - Prove that linear resolution does not p-simulate regular resolution.
4. **Johan Håstad**
   - Devise relations between monotone threshold circuits with bounded and unbounded weights. Non-monotone question is described in [7].

5. **Alexander Kulikov**
   - A function $f : \{0,1\}^n \to \{0,1\}$ is called an affine disperser for dimension d, if for every affine subspace $S \subseteq \{0,1\}^n$ of dimension at least d, f is not constant on S. This means that $n-d$ linear substitutions of the form $x_i = \bigoplus_{j \neq i} x_j \cdot b_j \oplus b_0$, where $b_i \in \{0,1\}$ do not make the function constant.
     Ben-Sasson and Kopparty, Shaltiel showed that there are affine dispersers for dimension $o(n)$ in P.
     Let us consider the following extension of affine dispersers. Now we allow linear and 'quadratic' substitutions. We start with a function of n variables. Then we make a substitution of the form $x_i = \bigoplus_{j \neq i} x_j \cdot b_j \oplus b_0$ or $x_i = (x_j \oplus b_j) \cdot (x_k \oplus b_k) \oplus b$, s.t. the substitution makes it a function of $n-1$ variables (i.e., after substituting x_i, it will never appear in the subsequent substitutions). We make $n-k$ substitutions as above and require the resulting function of k variables to be non-constant. Using a probabilistic argument one can show that these functions exist for dimension $k = o(n)$. My main question is whether it is possible to find dispersers of this kind for dimension $o(n)$ in NP?
   - Let $C(AND, OR, XOR)$ denote the circuit complexity (over the full binary basis B_2) of a function $f : \{0,1\}^n \to \{0,1\}^3$, such that $f(x) = (AND(x), OR(x), XOR(x))$.
     It is known that $2n-2 \leq C(AND, OR, XOR) \leq 2.5n$. Is it possible to improve the lower bound?
6. **Massimo Lauria**
   - There is a natural way to express in CNF form that a graph $G = ([n], E)$ contains a clique of size k (i.e., a set of k vertices pairwise connected by edges).
     If G has no k-clique then the corresponding CNF formula has a refutation. Furthermore, most algorithms to detect cliques in graphs would implicitly produce a resolution refutation of the k-clique formula, when they look for a k-clique in a graph that does not have any. The length of the refutation is proportional to the running time of the algorithm.
     For this reason it is interesting to determine how long is a refutation the k-clique formula: $n^{O(k)}$ is an obvious upper bound. Is this tight? Does the k-clique formula require a resolution refutation of size $n^{\Omega(k)}$ for some graph family?
     The CNF formulation of the clique formulas as well as further background can be found in [1,2].
7. **Jochen Messner**
   - Is there a $\leq_m^p$-complete set among all sets with an optimal acceptor?
   - Does every set with an optimal acceptor have a p-optimal proof system?
   - Is there a set outside P that has a p-optimal proof system?
     Some background information can be found in [4,6].
8. **Hunter Monroe**
   - Can hard instances be generated in various settings (hard tautologies to prove or to accept, hard inputs $< N, x, 1^t >$ to the complement of bounded halting) and given that such a construction would imply P$\neq$ NP, how could it circumvent the limits on diagonalization identified by Baker, Gill, and Solovay?
9. **Sebastian Müller**
   - In Parity Games you can easily construct gadgets that, when adjoined to any game graph, make the associated Parity Game trivial, but alteration of one specific edge, vertex or priority makes the gadget useless and therefore the game on the graph with the altered gadget is as hard to solve as the original one.

As these gadgets can be constructed for most classes of graphs (planar is a weak exception), it shows that most classes of game graphs over which Parity Games are feasible are not closed under the above alterations.

What happens if we are concerned with random edges, vertices or priorities? Can we construct a graph, where random alterations already lead to problems? Can we posibly add this to an exisiting graph and infer something in the light of what I said above? Also, what happens if we look at specific or random alterations on the random graph (perceived as a game graph)?

Background information on these problems can be found in [9].

10. **Rahul Santhanam**
    - For a deterministic Turing machine M which halts on all inputs, let $T_M(n)$ be the worst-case time complexity of M on inputs of length n. Consider the following 'running time estimation' problem: given n in unary, compute $T_M(n)$. Is there an exponential time-bounded machine M such that a polynomial-time solution to the running time estimation problem for M has interesting complexity-theoretic consequences, eg., a collapse of complexity classes?

11. **Alexander Smal**
    - The following are equivalent
      a. There is an optimal propositional proof system.
      b. TAUT has an almost optimal nondeterministic algorithm.
      c. There is a nondeterministic algorithm that decides p-Halt$_>$ problem.
      (Input of p-Halt$_>$ is a pair of nondeterministic Turing machine M and natural number n in unary. The problem is "does every accepting run of M on the empty input take more than n steps?")

    What is a heuristic analogue of this statement?

    Some background information can be found in [3, 8].

References

1 Olaf Beyersdorff, Nicola Galesi, and Massimo Lauria. Parameterized complexity of DPLL search procedures. *ACM Transactions on Computational Logic*, 14(3), 2013.

2 Olaf Beyersdorff, Nicola Galesi, Massimo Lauria, and Alexander Razborov. Parameterized bounded-depth Frege is not optimal. *ACM Transactions on Computation Theory*, 4(3), 2012.

3 Yijia Chen, Jörg Flum. A parameterized halting problem. *Lecture Notes in Computer Science*, 7370, 2012.

4 Johannes Köbler, Jochen Messner, Jacobo Torán. Optimal proof systems imply complete sets for promise classes. *Information and Computation*, 184(1), 2003.

5 Nicola Galesi, Pavel Pudlák, and Neil Thapen. The space complexity of cutting planes refutations. *Electronic Colloquium on Computational Complexity (ECCC)*, 21:138, 2014.

6 Jochen Messner. On optimal algorithms and optimal proof systems. In *Proc. STACS'99*, pages 541–550, 1999.

7 Mikael Goldmann, Johan Håstad and Alexander Razborov. Majority Gates vs. General Weighted Threshold Gates. *Journal of Computation Complexity*, 1(4), 1992.

8 Edward A. Hirsch, Dmitry Itsykson, Ivan Monakhov, Alexander Smal. On Optimal Heuristic Randomized Semidecision Procedures, with Applications to Proof Complexity and Cryptography. *Theory of Computing Systems*, 51(2), 2012.

9 Sebastian Müller. Graph structure and parity games. *Electronic Colloquium on Computational Complexity (ECCC)*, 21:121, 2014.

Participants

- Per Austrin
KTH Royal Institute of Technology, SE
- Olaf Beyersdorff
University of Leeds, GB
- Ilario Bonacina
University of Rome "La Sapienza", IT
- Samuel R. Buss
University of California – San Diego, US
- Igor Carboni Oliveira
Columbia Univ. – New York, US
- Ruiwen Chen
University of Edinburgh, GB
- Yijia Chen
Shanghai Jiao Tong Univ., CN
- Leroy Chew
University of Leeds, GB
- Andrew Drucker
University of Edinburgh, GB
- Susanna Figueiredo de Rezende
KTH Royal Institute of Technology, SE
- Jörg Flum
Universität Freiburg, DE
- Nicola Galesi
University of Rome "La Sapienza", IT
- Michal Garlik
Charles University – Prague, CZ
- Christian Glasser
Universität Würzburg, DE
- Andreas Goerdt
TU Chemnitz, DE
- Johan Hastad
KTH Royal Institute of Technology, SE
- Edward A. Hirsch
Steklov Institute – St. Petersburg, RU
- Dmitry Itsykson
Steklov Institute – St. Petersburg, RU
- Mikoláš Janota
INESC-ID – Lisboa, PT
- Emil Jerabek
Acad. of Sciences – Prague, CZ
- Alexander Knop
St. Petersburg State Univ., RU
- Johannes Köbler
HU Berlin, DE
- Jan Krajíček
Charles University – Prague, CZ
- Alexander S. Kulikov
Steklov Institute – St. Petersburg, RU
- Massimo Lauria
KTH Royal Institute of Technology, SE
- Barnaby Martin
Middlesex University, GB
- Jochen Messner
Ulm, DE
- Hunter Monroe
IMF – Washington, US
- Moritz Müller
Universität Wien, AT
- Sebastian Mueller
University of Toronto, CA
- Jakob Nordström
KTH Royal Institute of Technology, SE
- Jan Pich
Charles University – Prague, CZ
- Pavel Pudlák
Acad. of Sciences – Prague, CZ
- Benjamin Rossman
National Institute of Informatics – Tokyo, JP
- Zenon Sadowski
University of Bialystok, PL
- Rahul Santhanam
University of Edinburgh, GB
- Alan Selman
SUNY – Buffalo, US
- Alexander Smal
Steklov Institute – St. Petersburg, RU
- Dmitry Sokolov
Steklov Institute – St. Petersburg, RU
- David Steurer
Cornell University, US
- Jacobo Torán
Universität Ulm, DE

Report from Dagstuhl Seminar 14441

Modeling, Verification, and Control of Complex Systems for Energy Networks

Edited by

Alessandro Abate*[1], Martin Fränzle[2], Ian Hiskens[3], and Martin Střelec[4]

1 **University of Oxford, GB,** Alessandro.Abate@cs.ox.ac.uk
2 **Universität Oldenburg, DE,** martin.fraenzle@informatik.uni-oldenburg.de
3 **University of Michigan, US,** hiskens@umich.edu
4 **University of West Bohemia, Pilsen, CZ,** strelec@ntis.zcu.cz

Abstract

Power and energy networks are systems of great societal and economic relevance and impact, particularly given the recent growing emphasis on environmental issues and on sustainable substitutes (renewables) to traditional energy sources (coal, oil, nuclear). Power networks also represent systems of considerable engineering interest.

The aim of this Dagstuhl seminar has been to survey existing and explore novel formal frameworks for modeling, analysis and control of complex, large scale cyber-physical systems, with emphasis on applications in power networks.

Stochastic hybrid systems (SHS) stand for a mathematical framework that allows capturing the complex interactions between continuous dynamics, discrete dynamics, and probabilistic uncertainty. In the context of power networks, stochastic hybrid dynamics arises naturally: (i) continuous dynamics models the evolution of voltages, frequencies, etc.; (ii) discrete dynamics models controller logic and changes in network topology (unit commitment); and (iii) probability models the uncertainty about power demand, power supply from renewables and power market price.

The seminar has covered relevant approaches to modeling and analysis of stochastic hybrid dynamics, in the context of energy networks.

Seminar October 26–31, 2014 – http://www.dagstuhl.de/14441

1998 ACM Subject Classification B.2.2 Verification, D.2.4 Formal methods, Model checking, H.1.1 Systems theory, I.2.8 Control theory, I.6.1 Systems theory

Keywords and phrases Analysis, control, and verification of complex stochastic systems; formal synthesis; reliability engineering and assessment; energy networks

Digital Object Identifier 10.4230/DagRep.4.10.69

Edited in cooperation with Jan Krčál

* Thanks to the Lucy Halsall fund, Linacre College Oxford, for partial support.

Modeling, Verification, and Control of Complex Systems for Energy Networks, *Dagstuhl Reports*, Vol. 4, Issue 10, pp. 69–97

Editors: Alessandro Abate, Martin Fränzle, Ian Hiskens, and Martin Střelec

Dagstuhl Reports

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

1 Executive Summary

Alessandro Abate
Martin Střelec

The seminar has been focused on a number of selected topics from energy networks, with an emphasis on power systems that have great societal and economical relevance and impact. These represent systems of considerable engineering interest, since:

- they can be large-scale and can involve numbers of various devices interconnected in a complex manner.
- they are heterogeneous, that is they can be naturally modelled through a combination of continuous dynamical elements (to capture the evolution of quantities such as voltages, frequencies and generation output) and discrete dynamical components (to capture changes in the network topology, controller logic, state of breakers, isolation devices, transformer taps, etc.).
- they involve substantial stochastic components. Sources of uncertainty traditionally considered in power networks include hardware faults and unforeseen events, as well as stochasticity arising from continuous processes, particularly power demand. Furthermore, the increasing availability of renewable energy sources (e. g. photovoltaic panels, wind turbines, etc.) implies that uncertainty (for example, uncertainty in weather forecasts or cloud cover) also enters at the power supply side.
- some variables are only partially observable due to absence of real-time sensing circuitry in large parts of the existing power distribution network.

Reasonable and accurate analysis of future power networks needs models that seamlessly integrate behavioural patterns like complex interaction of continuous electrical phenomena (e. g. power flows) related to connected devices, discrete events caused by switching behaviour in circuitry, commitment of supplies and loads or by decisions of market participants, and the inherently stochastic behaviour of volatile supplies, demands and market prices.

In summary, the aim of the seminar has been to survey existing and explore novel formal frameworks for modelling, analysis and control of complex, large scale systems, with emphasis on applications in power networks. The seminar has hosted researchers and practitioners working on energy network application domains, in order to import related techniques for the study of energy grids in general, their analysis and energy management, which consists in control, coordination and dispatch of multiple generation, consumption and storage devices connected to the grid. Interactions among scientists and professionals from the heterogeneous research and application fields focused on power networks has highlighted opportunities for further research concerning expressiveness of models and scalability of the methods, as well as point to related efforts in the power network community.

General comments

The Seminar has run over the last week of October 2014 (27 to 31), has been well attended throughout the week, with about 40 participants. It has featured a fully packed program made up of presentations (at least 30), sustained discussions, and breakout sessions on three different topics. A final discussion session has concluded the proceedings of this event.

While the presence from academia has been preponderant, we have also been happy to see a number of active participants from the industry. The attendants expertise has been quite diverse. Academic participants have come with backgrounds in verification, control, and power systems. Alongside the participated and very open discussions, the seminar has additionally featured a hike and a dinner at a local restaurant.

Program

Talks have been categorised within the following clusters: Theory and Tools from Control; Theory and Tools from Verification; Topics in Power Networks; Smart/Micro Grids and Buildings.

Beyond these clusters, we have tried to diversify the program in order to optimally engage the audience. Discussions have been fostered via an afternoon breakout session, organised on Tuesday, the social activities on Wednesday afternoon, and the final session on Friday in the late morning.

There have been three breakout sessions, focusing respectively on

- modelling issues in energy/power systems;
- simulation issues in energy/power systems;
- demand response: control and verification.

The topics elaborated during the sessions are discussed in the ensuing sections, which report the notes that have come out of the discussions.

2 Table of Contents

3 Overview of Talks

3.1 Aggregation and Control of Populations of Thermostatically Controlled Loads by Formal Abstractions

Alessandro Abate (University of Oxford, GB)

Joint work of Abate, Alessandro; Sadegh Esmaeil Zadeh Soudjani
Main reference S. E. Z. Soudjani, A. Abate, "Aggregation and Control of Populations of Thermostatically Controlled Loads by Formal Abstractions," IEEE Transactions on Control Systems Technology, to appear.

This work discusses a two-step procedure, based on the use of formal abstractions, to generate a finite-space stochastic dynamical model as an aggregation of the continuous temperature dynamics of a homogeneous population of Thermostatically Controlled Loads (TCLs). The temperature of a TCL is described by a stochastic difference equation and the TCL status (ON, OFF) by a deterministic switching mechanism. The procedure is deemed to be formal, as it allows the quantification of the error introduced by the abstraction. As such, it builds and improves on a known, earlier approximation technique used in the literature. Further, the contribution discusses the extension to the instance of heterogeneous populations of TCLs by means of two approaches. It moreover investigates the problem of global (population-level) power reference tracking and load balancing for TCLs that are explicitly dependent on a control input. The procedure is tested on a case study and benchmarked against the mentioned existing approach in the literature.

References

1 S.E.Z. Soudjani and A. Abate, "Aggregation of Thermostatically Controlled Loads by Formal Abstractions," Proceedings of the 12th European Control Conference, pp. 4232–4237, Zürich (CH), 2013.
2 M. Kamgarpour, C. Ellen, S.E.Z. Soudjani, S. Gerwinn, J.L. Mathieu, N. Müllner, A. Abate, D.S. Callaway, M. Fränzle, and J. Lygeros, "Modeling Options for Demand Side Participation of Thermostatically Controlled Loads," IREP 2013, pp. 1–15.
3 S.E.Z. Soudjani and A. Abate, "Aggregation and Control of Populations of Thermostatically Controlled Loads by Formal Abstractions," IEEE Transactions on Control Systems Technology. 2014, In Print.

3.2 Spatial Temporal Logic Inference, Verification, and Synthesis

Calin A. Belta (Boston University, US)

Joint work of Kong, Zhaodan; Haghighi, Iman; Bartocci, Ezio; Aydin Gol, Ebru; Jones, Austin; Belta, Calin A.;
Main reference Z. Kong, A. Jones, A. Medina-Ayala, E. Aydin Gol, C. Belta, "Temporal Logic Inference for Classification and Prediction from Data," in Proc. of the 17th Int'l Conf. on Hybrid Systems: Computation and Control (HSCC'14), pp. 273–282, ACM, 2014.
URL http://dx.doi.org/10.1145/2562059.2562146

Networked dynamical systems are increasingly used as models for a variety of processes ranging from robotic teams to collections of genetically engineered living cells and power networks. As the complexity of these systems increases, so does the range of emergent properties that they exhibit. This presentation introduces a new logic called Spatial-Temporal Logic (SpaTeL) that is a unification of signal temporal logic (STL) and tree spatial superposition logic (TSSL). SpaTeL is capable of describing high-level spatial patterns that change over time, e. g., "Power consumption in the northwest quadrant of the city drops below 100 megawatts if the power consumption in the southwest quadrant remains above 200 megawatts for two hours." A statistical model checking procedure that evaluates the probability with which a networked system satisfies a SpaTeL formula is presented. A synthesis procedure that determines system parameters maximizing the average degree of satisfaction, a continuous measure that quantifies how strongly a system execution satisfies a given formula, is also introduced. The approach is illustrated on two systems: a biochemical reaction-diffusion system and a demand-side management system for a smart neighborhood.

3.3 Hybrid stochastic systems: simulation in Modelica

Marc Bouissou (Ecole Centrale Paris, FR)

Joint work of Bouissou, Marc; Elmqvist, Hilding; Otter, Martin; Benveniste, Albert
Main reference M. Bouissou, H. Elmqvist, M. Otter, A. Benveniste, "Efficient Monte Carlo simulation of stochastic hybrid systems", in Proc. of the 10th Int'l Modelica Conf., Linkþing Electronic Conference Proceedings, Vol. 96, pp. 715–725, Linköping University Electronic Press, 2014.
URL http://dx.doi.org/10.3384/ECP14096715

The purpose of this presentation is to show some of the results of the MODRIO project, relative to the practical use of hybrid, often stochastic, models of complex systems in the industry. This project is all based on Modelica models and tools. The general idea is to extend the use of the models usually built for the design of systems, to the exploitation phase of their lifecycle. This presentation is focused on the reuse of Modelica design models for performing dependability analysis. Depending on the degree of "hybridicity", several approaches can be used, going from enriching the Modelica deterministic model corresponding to nominal behavior by random features, to abstracting the model into a language dedicated to discrete systems, even Boolean abstractions like fault trees. In the presentation we explain an optimal algorithm for performing Monte Carlo simulation of a PDMP (piecewise deterministic Markov process). This algorithm was published in [1]. Then we show the needs for modeling hybrid stochastic systems in Modelica models. The main challenges (that are being addressed in the MODRIO project) are about multi-mode hybrid systems where the number of state variables changes when the system changes mode, and the introduction of stochastic transitions in continuous time state machines of Modelica. We will show the model and simulation results for a simple pedagogical example: a room heated by a heater subject to random failures and repairs.

References

1 "Efficient Monte Carlo simulation of stochastic hybrid systems". Marc Bouissou, Hilding Elmqvist, Martin Otter, Albert Benveniste. Modelica conference 2014.

3.4 Dynamic coupling, nonlinear output agreement and power network control

Claudio De Persis (University of Groningen, NL)

Joint work of Buerger, Mathias; De Persis, Claudio; Trip, Sebastian
Main reference M. Bürger, C. De Persis, "Dynamic coupling design for nonlinear output agreement and time-varying flow control," Automatica, 51(Jan. 2015):210–222, 2015; pre-print available as arXiv:1311.7562v1 [cs.SY].
URL http://dx.doi.org/10.1016/j.automatica.2014.10.081
URL http://arxiv.org/abs/1311.7562v1
Main reference S. Trip, M. Bürger, C. De Persis, "An internal model approach to frequency regulation in power grids," arXiv:1403.7019v2 [cs.SY], 2014.
URL http://arxiv.org/abs/1403.7019v2

We discuss recent results on the problem of output agreement in networks of nonlinear dynamical systems under time-varying disturbances. The key to these results is the use of dynamic diffusive couplings and internal model controllers. For the class of incrementally passive systems, constructive sufficient conditions for output agreement are presented. The proposed approach lends itself to solve flow control problems in distribution networks. As a case study we show how recently proposed controllers for frequency regulation can be interpreted as internal model controllers and discuss how the approach can shed new light to the problem.

3.5 Plug-and-Play Control and Optimization in Microgrids

Florian Dörfler (ETH Zürich, CH)

Main reference F. Dörfler, J. W. Simpson-Porco, F. Bullo, "Plug-and-Play Control and Optimization in Microgrids," in Proc. of the 2014 IEEE 53rd Conf. on Decision and Control (CDC'14), to appear; pre-print available from author's webpage.
URL http://motion.me.ucsb.edu/pdf/2013y-dsb-conference.pdf

Microgrids are low-voltage electrical distribution networks, heterogeneously composed of distributed generation, storage, load, and managed autonomously from the larger transmission network. Modeled after the hierarchical control architecture of power transmission systems, a layering of primary, secondary, and tertiary control has become the standard operation paradigm for microgrids. Despite this superficial similarity, the control objectives in microgrids across these three layers are varied and ambitious, and they must be achieved while allowing for robust plug-and-play operation and maximal flexibility, without hierarchical decision making and time-scale separations. In this seminar, we explore control strategies for these three layers and illuminate some possibly-unexpected connections and dependencies among them. We build upon a first-principle model and different decentralized primary control strategies such as droop, quadratic droop, and virtual oscillator control. We motivate the need for additional secondary regulation and study centralized, decentralized, and distributed secondary control architectures. We find that averaging-based distributed controllers using communication among the generation units offer the best combination of flexibility and performance. We further leverage these results to study constrained AC economic dispatch in a tertiary control layer. Surprisingly, we show that the minimizers of the economic dispatch optimization problem are in one-to-one correspondence with the set of steady-states reachable by droop control. This equivalence results in simple guidelines to select the droop coefficients, which include the known criteria for power sharing. Finally, we illustrate the performance and robustness of our designs through through hardware experiments.

3.6 Multi-Objective Parameter Synthesis in Probabilistic Hybrid Systems

Martin Fränzle (Universität Oldenburg, DE)

Joint work of Abate, Alessandro; Fränzle, Martin; Gerwinn, Sebastian; Katoen, Joost-Pieter; Kröger, Paul

Technical systems interacting with (parts of) the real world can be modelled elegantly using probabilistic hybrid automata (PHA). Parametric probabilistic hybrid automata are dynamical systems featuring hybrid discrete-continuous state dynamics and parametric probabilistic branching, thereby extending PHA by capturing a family of PHA in a single model. Such systems have a broad range of application from controller synthesis to network protocols. In our talk, we presented a novel method to synthesize parameter instances (if such exist) of such models that satisfy a multi-objective bounded horizon specification over expected rewards. Our technique combines three distinct techniques: (1.) statistical model checking of a model instance devoid of parameters, (2.) a symbolic version of importance sampling yielding an arithmetic constraint reflecting the parameter dependence of the expected rewards, and (3.) SAT-modulo-theory (SMT) solving applied to that constraint for finding a feasible parameter instance consistent with the multi-objective design goal modulo sampling effects. Using a check-and-refine approach enhancing the constraint based on counterexamples, we are able to provide strong statistical guarantees on feasibility of the synthesized parameter instance.

3.7 Hybrid Systems: How Nonlinear Can We Go?

Sicun Gao (Carnegie Mellon University – Pittsburgh, US)

I will present the framework of delta-complete analysis for reachability problems of a very general class of nonlinear hybrid systems. We perform bounded reachability checking or invariant-based reasoning through solving delta-decision problems over the reals. The techniques take into account of robustness properties of the systems under numerical perturbations. The verification problems become much more mathematically tractable in this new framework. I will demo our open-source tool dReach, which scales well on several highly nonlinear hybrid system models that arise in biomedical and robotics applications.

3.8 Resilience modelling for smart neighborhoods

Boudewijn Haverkort (University of Twente, NL)

The use of renewable energy in houses and neighborhoods is very much governed by national legislation and has recently led to enormous changes in the energy market and poses a serious thread to the stability of the grid at peak production times. One of the approaches towards a more balanced grid is, e. g., taken by the German government by subsidizing local storage for solar power.

While the main interest of the energy operator and the government is to balance the grid and still have positive revenues, thereby ensuring grid stability, the main interest of the client is twofold: the total cost for electricity should be as low as possible and the power supply be as resilient as possible in the presence of power/grid outages. Clearly, these two objectives highly depend on the availability and capacity of local storage, and the overall strategy followed by the local controller.

We present a Hybrid Petri net model of a house that is mainly powered by solar energy with a local storage unit and subsequently analyze the impact of different storage strategies on the resilience of the power supply and the overall cost of electricity, for different production and consumption patterns.

3.9 Dynamics of Generic Wind Turbine Generator Models

Ian Hiskens (University of Michigan – Ann Arbor, US)

Main reference I. A. Hiskens, "Dynamics of type-3 wind turbine generator models," IEEE Transactions on Power Systems, Vol. 27(1):465–474, 2012.
URL http://dx.doi.org/10.1109/TPWRS.2011.2161347

The talk provides an analysis of a generic model describing the dynamic behaviour of type-3 wind turbine generators (WTGs). The behaviour of this model is governed by interactions between the continuous dynamics of state variables and discrete events associated with controls. It is shown that these interactions can be quite complex, and may lead to switching deadlock that prevents continuation of the trajectory. Switching hysteresis is proposed for eliminating deadlock situations. Various type-3 WTG models include control blocks that duplicate integrators. It is shown that this leads to non-uniqueness in the conditions governing steady- state, and may result in pre- and post-disturbance equilibria not coinciding.

3.10 Diagnostics and maintenance of HVAC systems

Ondřej Holub (Honeywell Prague Laboratories, CZ)

Joint work of Abate, Alessandro; Adzkiya, Dieky; Berka, Jan; Endel, Petr; Holub, Ondřej

Buildings consume more than 40% of energy in Europe. Smart building automation systems can considerably contribute to the balance in the grid at a neighborhood or district scale. In order to sustainably achieve this goal a continuous commissioning and maintenance of building automation systems is needed. Building maintenance typically includes many tasks solved by algorithms that are designed using different methodologies. Integration of the dedicated algorithms into one solution represents an ever more complex effort. A generic approach to the HVAC maintenance is presented in the talk, which solves this problem by formulating it in the framework of Stochastic Hybrid Systems.

3.11 Consensus + innovations based distributed optimization in electric power systems

Gabriela Hug-Glanzmann (Carnegie Mellon University, US)

Joint work of Hug-Glanzmann, Gabriela; Kar, Soummya; Mohammadi, Javad
Main reference J. Mohammadi, G. Hug, S. Kar, "Role of Communication on the Convergence Rate of Fully Distributed DC Optimal Power Flow," in Proc. of the 2014 IEEE Int'l Conf. on Smart Grid Communication (SmartGridComm'14), pp. 43–48, IEEE, 2014.
URL http://dx.doi.org/10.1109/SmartGridComm.2014.7007620

Traditionally, electric power is generated in bulk power plants and transmitted over a power grid with limited control capabilities to supply inflexible loads. With the trend moving towards more distributed generation and storage resources, flexible demand and increased power flow control capabilities, the number of control variables in the system increases significantly. For the coordination of the settings of these control variables, a more distributed control architecture may become more suitable than the traditional centralized approach. Hence, in this talk, the formulation of consensus + innovations based distributed optimization approaches applied to various optimization problems in power systems is discussed and simulation results which serve as a proof of concept are presented.

3.12 Provisioning of Energy Flexibility between Supply and Demand

George B. Huitema (University of Groningen, NL)

Currently we are facing the transition of the traditional Energy world to a Smart Energy System. Flexibility is key in Smart Energy. Flexibility here is not about the flow of energy itself but about the willingness or ability of energy units (i. e. households, generators, battery elements) to vary respectively in their consumption, production or storage. The variation thus may concern parameters like time (shifting loads), quantity, quality or e. g. type of energy (hybrid: electricity, heat, gas). Many Smart Grid business models, like Demand-response programs, deal with selling or buying flexibility. This is driven by revenues, savings or incentives. Related to the business relationships we have to take care of the corresponding (Smart Grid) Billing. No Billing is Killing. In this presentation I focus on the next step in the transition to Smart Energy Systems, that is not work anymore on implementation of single technologies but work on "Big Numbers" (1k–1M), and thus on complexity, i. e. system integration, aggregated flexibility and large scale business processes. In particular I describe the latest developments in

1. Smart Grid Coordination (2nd generation multi-stakeholder transactive energy algorithm for multi-objective optimization, example is the TNO PowerMatcher agent-based Energy Management System);
2. Active Customer Participation;
3. Flexible Power Architecture Interface (FPAI);
4. Hybrid System Integration (HESI) and
5. Smart Grid Billing (or SmarteBility).

Finally, I present the case of a Flexibility Provider, i. e. an aggregator who services to other parties the flexibility contracted with consumers. As a statement, I assume that the Flexibility Provider is a spin-doctor in this business proposition.

3.13 Thermal Loads for Ancillary Services

Maryam Kamgarpour (ETH Zürich, CH)

The increase in fluctuating weather-based renewable generation increases the need for the so-called ancillary services, which help in balance of supply and demand. Control reserves are the most important form of ancillary services and are today mainly covered by conventional generators, such as hydro power plants. I will discuss the use of thermal loads as additional means for ancillary services. Two main options for thermal loads will be discussed: first, control of Heating, Ventilation, and Air Conditioning (HVAC) systems of an aggregate of several office buildings; second, control of a large number of household appliances. I will discuss our modeling, optimization, and control tools developed to quantify the potential of these loads to serve in the ancillary service market.

3.14 Sooner is Safer than Later in Branching Time too

Joost-Pieter Katoen (RWTH Aachen, DE)

Joint work of Hermanns, Holger; Katoen, Joost-Pieter; Song, Lei; Zhang, Lijun

I present a formal characterization of branching-time safety and liveness properties for timed systems. As in the classical branching-time setting, I distinguish universally and existentially safe (and live) properties. I characterize these property classes topologically as well as by sound and complete fragments of timed CTL. For finitely branching timed properties, we present an algorithm that decomposes an arbitrary timed CTL formula into an equivalent conjunction of a safety and a liveness property, both represented by timed alternating tree automata.

3.15 Recharging Probably Keeps Batteries Alive

Jan Krčál (Universität des Saarlandes, DE)

Joint work of Hermanns, Holger; Krčál, Jan; Nies, Gilles

The kinetic battery model is a popular model of the dynamic behavior of a conventional battery, useful to predict or optimize the time until battery depletion. The model however lacks certain obvious aspects of batteries in-the-wild, especially with respect to (i) the effects of random influences and (ii) the behavior when charging up to capacity bounds.

This paper considers the kinetic battery model with bounded capacity in the context of piecewise constant yet random charging and discharging. The resulting model enables the time-dependent evaluation of the risk of battery depletion. This is exemplified in a power dependability study of a nano satellite mission.

3.16 Mosaik – a Framework for Smart Grid Co-Simulation

Sebastian Lehnhoff (OFFIS – Oldenburg, DE)

Main reference S. Rohjans, S. Lehnhoff, S. Schütte, F. Andrén, T. Strasser, "Requirements for Smart Grid Simulation Tools," in Proc. of the 23rd IEEE Int'l Symp. on Industrial Electronics (ISIE'14), pp. 1730–1736, IEEE, 2013.
URL http://dx.doi.org/10.1109/ISIE.2014.6864876
URL https://mosaik.offis.de

Future energy systems will have to integrate large varieties and numbers of active components (generation and consumption devices as well as operational equipment) into energy management or operating schemes allowing for an automated on-line optimization of appropriate systems. Although the necessary Information and Communication Technologies (ICT) and automation technologies are well known and established in other application domains, their implementation and prolonged use in (safety-critical) future Smart Grids is hitherto untested and thus fraught with potential risk for relevant stakeholders in the energy domain, e. g., Distribution and Transmission System Operators. Therefore, novel ICT-based systems e. g., for control and protection systems, adaptive ancillary service provision through decentralized units have to be tested rigorously in advance in order to minimize uncertainties in their prospective adoption.

When addressing the key question of how and on what parts of the system to test ICT components, the scope of future Smart Grids is hard to determine. Future energy systems will depend not only on unit (generation and consumption) commitment and on the utilization of the operating equipment in between but also on weather phenomena, user preferences or strategies, the utilization and thus availability of the underlying ICT system, market prices, regulatory constraints and governance to name but a few of the relevant facets. When taking complex interactions of these facets into account matters become even more complex. Omitting allegedly irrelevant facets might prove dangerous as even small effects quickly gain relevance through scaling a prominent example was the European 50.2Hz frequency problem. Formal analysis of such a system is not feasible anymore and realistic field tests or experimental hardware environments representative of the complex and interactive system too expensive. This paves the way for couplings of real hardware interacting with simulated environments (hardware-in-the-loop) or even pure software-based simulations in power system design and analysis. However, there is a need for proper simulation environments supporting the flexible integration of (oftentimes black- box-)models of various origin and representation (including hardware) into a functional environment.

In this talk we will present mosaik – a flexible Smart Grid co-simulation framework for reusing and coupling existing simulation models and simulators to create large-scale Smart Grid scenarios yielding thousands of simulated entities distributed across multiple simulator processes. These scenarios may then serve as a test bed for various types of control strategies (e. g., Multi-Agent Systems (MAS) or centralized control) in complex an realistic scenarios. Following mosaiks scenario design we will specify the state-of-the-art, identify current challenges with respect to modeling, simulation and uncertainty quantification in Smart Grid design and analysis.

3.17 Modeling, Validation, and Calibration of Power System Models.

Bernard Lesieutre (University of Wisconsin – Madison, US)

In this seminar we present results on power plant model validation. Using PMU measurements located at the point of interconnection, power plant models may be validated/invalidated. Simulations using dynamics models are compared to data from disturbances. When the models fail to match the measurements, the errors can be used to recalibrate the models. Issues involve the detection of discrete changes including control loop failures, and the non-uniqueness of parameter profiles to match data. Three illustrative examples are provided.

3.18 Mean field based decentralized control of power system components for a smoother integration of renewable sources

Roland P. Malhame (Polytechnique Montreal, CA)

The past years have seen a significant increase in the level of renewable energy sources in the power generation mix, and a number of countries have set for themselves ambitious goals in this respect. However, the intermittent character of most of the renewable sources creates particular challenges in terms of maintaining the balance between generation and load. In this context, energy storage becomes an important tool for making higher renewable penetration levels possible. Much energy storage of electrical origin is present in the power system; it is associated with temperature sensitive devices such as air conditioners, electric space heaters,electric water heaters, and refrigerators; also electric vehicle batteries could collectively represent an important energy storage capability. However besides the need for a communication link between the generation authority and the loads of interest, as well as the challenge of enlisting the agreement of customers, an important difficulty in tapping into the energy storage, remains the sheer number (in the millions) of devices which would have to be reached and properly coordinated for this purpose. Mean field control theory appears to be the right tool for dealing with such a challenge as it turns the complexity of large numbers into an ally, by harnessing the power of the law of large numbers to achieve predictability of aggregate load behaviors. It is a blend of the mathematical theory of Games with the Physics theory of Statistical Mechanics. We present a hierarchical control architecture with an upper scheduling level working with aggregate models of the devices to be controlled. Using all relevant macroscopic information, the upper level uses mathematical programming to calculate optimal mean power or energy trajectories for the average individual device in a given class. It is left up to a lower level mean field controller to synthesize the decentralized feedback control laws which will lead the group to track the desired aggregate behavior. The basics of the theory and simulation results are presented.

3.19 Learning and data based optimization with application to generation and load side control of uncertain power systems

Kostas Margellos (University of California – Berkeley, US)

Joint work of Margellos, Kostas; Prandini, Maria; Lygeros, John; Oren, Shmuel
Main reference K. Margellos, M. Prandini, J. Lygeros, "On the connection between compression learning and scenario based optimization," IEEE Transactions on Automatic Control, to appear; pre-print available as arXiv:1403.0950v2 [cs.SY].
URL http://arxiv.org/abs/1403.0950v2

Making optimal decisions in an uncertain environment is a challenging task. In the same time, advances in many engineering disciplines have led to a huge amount of easily accessible data. This big data trend, leads to a paradigm shift in control and optimization, rendering data driven control an alternative to deterministic or robust techniques. We first focus on the problem of designing a decision policy that optimizes a certain criterion and is immunized against data uncertainty. We show that the resulting decision can be accompanied with a performance certificate that is provided a-priori to the decision maker and encodes the confidence with which the decision maintains its robustness properties against uncertainty realizations other than those included in the data. We relate the issue of certificate provision with learning and generalization paradigms in machine learning and analyze the implications of this approach to optimization problems with certain structural characteristics. We next show how the developed machinery can be applied to different generation and load side control problems in power systems with renewable energy sources. We focus on the problem of stochastic unit commitment and on a novel paradigm and pricing analysis for demand side management. The proposed framework allows us to provide certificates regarding the performance of the underlying system and is in contrast to earlier approaches to such problems, that are typically restricted to rule based or ad-hoc methodologies.

3.20 Uncertain power system reserves from loads

Johanna Mathieu (University of Michigan – Ann Arbor, US)

Joint work of Mathieu, Johanna; Vrakopoulou, Maria; Andersson, Göran; Li, Bowen; Shen, Siqian; Zhang, Yiling
Main reference M. Vrakopoulou, J. L. Mathieu, G. Andersson,"Stochastic Optimal Power Flow with Uncertain Reserves from Demand Response," in Proc. of the 2014 47th Hawaii Int'l Conf. on System Sciences (HICSS'14), pp. 2353–2362, IEEE, 2014.
URL http://dx.doi.org/10.1109/HICSS.2014.296

Aggregations of electric loads can provide reserves (i. e. back-up capacity) to power systems, helping the system to manage real-time supply-demand mismatch. However, the amount of reserve capacity available from loads is a function of a variety of stochastic factors including weather and load usage patterns. In this talk, I will describe how aggregations of thermostatically controlled loads (each individually represented as a stochastic hybrid system) can be modeled as thermal energy storage. Using these models, we can tackle both control problems (e. g., energy price arbitrage) and planning problems (e. g., stochastic unit commitment). For the latter, I will describe recent work that proposed "uncertain reserves" where the possible reserve capacity is not known a priori. In that case, the system operator solves a stochastic optimization problem to manage uncertainty stemming from renewable generation, load consumption, and load control uncertainty. We explore a variety of methods to solve this problem including scenario-based methods, analytical reformulations that assume knowledge of uncertainty distributions, and distributionally robust optimization.

3.21 Electrical Transmission Grids: Needs and Challenges

Patrick Panciatici (RTE Energy – France, FR)

3.21.1 Needs from a Grid Operator Perspective

The electrical grids and their management become more and more complex. This state of affairs has different causes that will not disappear in the near future. In Europe, the first reason is the massive integration of renewable but generally intermittent generation in the system. Power flows in the grid are created by differences in the location of sinks and sources. With a significant amount of intermittent generation, the predictability of the sources (location and amount of power injections) decreases and affects the predictability of the flows. Furthermore, some of these new power plants could be small units (e. g. PV) connected to the distribution grid, changing the distribution grid into an active system. Moreover, Transmission System Operators (TSOs) have a poor observability of these power injections and have no control at all over them. Another factor is the inconsistency between the relatively short time to build new wind farms (2 or 3 years) and the time to go through all administrative procedures to build new lines (more than 5 years everywhere in Europe). In Europe, the best locations for wind farms are mostly along the coasts and offshore, while for photo-voltaic generation they are in the south of Europe. Since these locations do not generally match those of the large load centers, a transmission network is required and this network will have to cope with the variability of the flows induced by the stochastic nature of the novel generation subsystems. The second main reason is that it is more difficult than ever to build new overhead lines because of low public acceptance and "Not In My BackYard" (NIMBY) attitude. People are more and more afraid of hypothetical electromagnetic effects or just don't like to see big towers in the landscape and in particular in protected areas which are more and more numerous around Europe. It is very difficult to explain the need for new power lines to people who already have access to electricity at a reasonable price and with high reliability. An increase in the European Social Welfare with a positive feedback for the European economy and hopefully for all European citizens is a concept that is too theoretical compared to the negative local impact. Alternative solutions are technically complex, costly and need more time to be deployed. The third reason is linked to the setup of electricity markets crossing the administrative and historical borders. Generators, retailers and consumers view the transmission system as a public resource to which they should have unlimited access. This approach has the desirable effect of pushing the system towards maximization of the social welfare and an optimal utilization of the assets. However, this optimization is constrained by security considerations because widespread service interruptions spanning over long periods of time are unacceptable in our modern societies due to their huge economic and social costs. Since TSOs are responsible for maintaining the reliability of the electric power system, they must therefore define the operating limits that must be respected. As in any constrained optimization problem, the optimal solution towards which the market evolves tends to be limited by these security constraints. The stakeholders therefore perceive reliability management by the TSOs as constraining their activities and reducing the European Social Welfare rather than as enablers of this large physical market place, as it would be the case if the grid were a copper plate. The transparent definition and the precise assessment of the distance to these limits thus become more and more critical. The last reason is that the aging of grid assets needs increasing attention. A significant part of the European grids' assets are more than 50 years old. Asset management and maintenance in systems that can't be stopped, are extremely challenging and need to be precisely anticipated when large numbers of assets are approaching simultaneously the end of their expected life times. To maintain the security of the supply in this context, TSOs have to change the architecture of the system by considering the following technologies:

- Long distance HVAC underground cables with large reactive compensators.
- HVDC underground cables in parallel with the AC grid with smart controls of AC/DC converters.
- And, ultimately, HVDC grids, first to connect efficiently offshore wind farms and then to provide cheaper interconnections between distant areas.

Meanwhile, TSOs will try to optimize the existing systems by adding more and more special devices such as Phase Shifting Transformers, Static VAr Compensators and advanced controls and protection schemes, taking also advantage of the flexibility provided by HVDC links embedded in AC grids. At the same time, demand response or dispersed storage could offer new ways to control the system, even if business models and costs are still questionable. But in any case, this flexibility will require a rethinking of historical operating practices where grid operators made the assumption that the load is an uncontrollable exogenous stochastic variable. We have heard so often in conferences, seminars and workshops, that the power grid will soon be operated very near to its limits, so that this statement has become a cliché. This cliché is now a reality. To be more precise, it is no longer possible to respect the classical preventive N° 1 security standards during all hours in a year. The system is indeed no longer able to survive all single faults without post-fault actions, i. e. corrective controls. More and more corrective control strategies are hence elaborated and prepared to maintain the security of the system. The number of hours during which the system requires corrective actions to be secure is increasing, and that seems to be a natural trend associated with the massive integration of intermittent generation. More and more local or centralized Special Protection Schemes (SPS)/Remedial Actions Schemes(RAS) are deployed to implement automatically some of these corrective actions based on advanced measurement devices (Phasor Measurement Units, Dynamic Line Ratings, ...) and high bandwidth communication networks. Grid operators have to manage an extremely complex decision making process in order to ensure the reliability and quality of supply at minimal cost over different time horizons. For the sake of clarity, while not aiming at being exhaustive, the following problems need to be dressed by the grid operators:

- Long term (10–20 years): planning stage – where to build new power lines? which technology? which capacity?
- Mid term (2–5 years):
  - installation of control devices: substation design, var/reactive support, PSTs, replacement of conductors,SPS/RAS design;
  - asset management and maintenance: which equipment to upgrade, to replace, to repair and when?
- Short term (monthly–weekly): outage management, must-run generators, preparation of corrective actions, required margins.
- Real Time (two days ahead to real time):
  - interaction with energy markets: definition of grid capacities;
  - selection of substation's topology, settings of SPS/RAS, adjustment of generating units.

In all these contexts, the grid operators want to make "optimal" decisions over these different time horizons, even if some decision making processes are currently not formalized mathematically as optimization problems but are rather based solely on knowledge of experts. However, as complexity increases, decision support tools become mandatory to help these experts to make their decisions:

- For the long term planning, there is hyper uncertainty associated with the implementation of energy transition policies and long term market behavior (Priority to renewable energies, Demand Growth in context of efficiency promotion, Technology Costs: electrical batteries, Demand Response, EV, Distributed Generation, Carbon Tax, Fuel Costs) and Grid operators have to make robust decisions based on multi-scenario grid planning.
- In all these processes, the increasing level of uncertainty associated to wind and solar power must be taken into account, hence pushing towards the use of probabilistic methods.
- Operation nearest to the limits requires an accurate modeling of all pieces of equipment, of the corrective actions and of the dynamic behaviors, so as to allow an accurate enough assessment of security margins. Moreover, the active constraints could be related to voltage/reactive or stability issues and not only to thermal limits.

Grid operators must ensure an adequate consistency between these decision making processes. They are in fact multistage decision making processes considering all the different time horizons. At the planning stage, they have to consider the decisions which could be made in lower level problems: asset management and operation and the same between asset management and operation. The modeling of these lower level problems seems very challenging when these lower level problems become more complex. Approximations are required and relevant "proxies" must be found for this modeling. In this paper we defend the idea that in order to address all these different questions, it is valuable to explicitly formulate them as optimization problems. Most of these problems, once stated, are hard to solve exactly. On the other hand significant progress has been collected in the recent years both in computational and in mathematical respects. The goal of the paper is to highlight the main avenues of progress in these respects and explain how they can be leveraged for improving power systems management.

3.21.2 Taxonomy of Optimization Problems

The objective of this section is to dig into the different ways one can formulate and model optimization problems to be addressed in power systems (with a focus on system-wide problems, as those addressed by TSOs).

A. Modeling the optimization problem from a formal viewpoint. Based on examples, we discuss the intrinsic nature of the different optimization problems, by distinguishing different possible formulations. The general formulation can be summarized as a multistage decision making problem under uncertainty. But the formulation of this very general problem depends on the different time horizons and the type of decisions to make. We can divide these decisions in three classes, illustrated here by an analogy with IT systems. 1) Decisions changing the structure of the system (developing the hardware) 2) Decisions changing policies or control/protection schemes (developing the software) 3) Decisions modifying the operating points of the system (selecting input data to run the software on the hardware) We focus our discussion on the first problem which is the most challenging. This problem ideally requires the modeling of all the aspects of power systems: from possible long term energy policies to system operation using realistic modeling of the physical system and expectations of the grid users. The decisions related to the structure of the system

(”hardware”) take long time to be implemented, they have to go through long permitting processes and need quite long construction times. They are investment decisions. The objective is to optimize the associated capital expenditures (capex) by comparing them to future operational expenditures (opex) saving. The time frame is varying from around ten years for the construction a new power line to one year for changing conductors on existing power lines. A part of the problem is to choose the relevant mixture of technological options: ac overhead power lines, ac underground cables, hvdc links, phase shifter transformers (PSTs), new conductors for existing power lines, new reactive compensation devices, ... The problem can be formulated as stochastic dynamic programming problem. For long term expansion planning, scenario-based approaches seem the most attractive formulations in order to ensure some level of robustness as proposed in [1]. How to define reliability criteria and how to implement them, are key questions in these optimization problems. ”Energy Not Served” or ”Loss of Load” are generally used. An “artificial” monetization is performed and estimated costs are associated to these indexes. These large costs are simply added to operational expenditures. In stochastic formulations, generally only expected values are minimized without any cap on the maximum risk. This could be questionable and chance constraint programming or robust optimization could offer more relevant solutions. We could imagine that a generalization of Demand Response could change dramatically the definition of reliability and the foundations of power system design,pushing to less ”hardware” and more ”software” solutions as anticipated very optimistically in 1978 by F. C. Schweppe [2]. A review of the current formulation and associated optimization problem is mandatory as proposed for example in the two on-going European projects: e-HIGHWAY2050 [3] and GARPUR [4]. In this global optimization problem, the sub problem on selection of relevant technological options leads to a combinatorial optimization very similar to a ”knapsack” problem, increasing even more the complexity. We can identify three different dimensions: spatial, temporal and stochastic. The spatial complexity is increasing: ”more and more the electrical phenomena don’t stop at administrative borders”. We have to consider systems very extended (Pan- European Transmission System, Eastern or Western Interconnection in US, . . .) and at the same time local active distribution grids. Time constants range from milliseconds to several years, leading to temporal complexity. Uncontrollable load and renewable energy sources implies to take into account more than ever stochastic behaviors. Considering spatial, temporal and stochastic complexity all together is still out of reach.Trade-offs must be made to take into account at most two of them in details at the same time and using approximation for third one. The appropriate modeling of uncertainties is also a key factor to find realistic optimal solutions. The spatial and temporal correlations between these uncertainties must be taken into account not to be too optimistic or too pessimistic. This pushes towards probabilistic methods and risk based approaches. When the probabilistic properties of the uncertainties are only partially known, generalized semi-infinite programming seems an appealing method proposing to find robust solutions when the uncertainties live in a defined domain. The objective function is related to satisfaction of the grid users (consumers and suppliers): maximization of social welfare. We need to estimate the expectations and the behaviors of the grid users. For long term decisions, it seems reasonable to simulate a “perfect market” leading to a global rational behavior minimizing the costs. For more short term decisions, it could be important to simulate the actual behavior of the market players and the imperfect market design. These estimations could be formulated as optimization problems based on game theory finding Nash equilibrium.In practice, we could have to formulate multi-objective optimization problems which are generally transformed in a single optimization problem using a weighted sum of the different objective functions

$\min(w1.f1(x) + w2.f2(x) + +wn.fn(x))$: (For example, we want to minimize the production costs and the amount of emitted CO_2). Finding the associated weighting factors could be difficult and questionable. A more rational approach should be to formulate a true multi-objective function $\min(f1(x); f2(x); ; fn(x))$: But for a nontrivial multi-objective optimization problem, there does not exist a single solution that simultaneously optimizes each objective. In that case, the objective functions are said to be conflicting, and there exists a (possibly infinite number of) Pareto optimal solutions. A solution is called non dominated, Pareto optimal, Pareto efficient or non inferior, if none of the objective functions can be improved in value without degrading some of the other objective values. This leads to complex optimization problems which could be solved using meta-heuristics methods. We could see that power system management could lead to a large diversity of optimization problems. The proper formulation of each problem has to be well thought out before searching for computational solutions.

Modeling the physics of the power system. The objective of this section is to analyze the physics and technological constraints arising from the power system, and explain their implications in terms of the nature of the above formulated optimization problems. The quality of physical modeling of power systems used in optimization problems is essential in order to make "optimal" decisions. Solving optimization problems with a high accuracy based on not realistic enough modeling is useless. We need to find the right balance between realism and complexity. The usage of static and deterministic modeling using a linearization of the associated mathematical formulations should be questioned in the new context presented in the introduction. A significant number of controls in electrical grids are discrete: switch on/off of breakers, switch on/off capacitor or reactor banks, tap changers on transformers, generating units producing with non zero minimal active power when they are started. These controls become integer variables in optimization problems and their treatments require a special attention; with a naive relaxation (round off strategy) is not always possible even to find feasible solutions. Some controls and protection schemes implemented in local or centralized SPS/RAS are not event-based but measurement based. They acts conditionally when a measurement or a set of measurements don't fulfill a given rule (for example, when a measurement value is beyond a given limit). This kind of behavior must be taken into account in optimization problem. This leads to conditional corrective actions and the modeling of this type of hybrid system (continuous and discrete) requires binary variable [5]. Some active constraints in power systems could be more and more related to stability and the system dynamic behavior. The ultimate solution should be to use a DAE-constrained optimization formulation but a reasonable first step could be to use "proxies" to hide this complexity. The idea is to learn using Monte-Carlo simulation, rules which ensure that the study case has a very low probability to be prone to stability issues [6] and to introduce these rules in static optimization problems. Power system planning and operation raises many important decision making problems, which can generally be stated as large-scale, non-linear, mixed-integer continuous, non-convex, stochastic and/or robust optimization problems. In the last years, many progresses have been made in the theory and implementation of optimization algorithms, driven by research in applied mathematics and by multitudinous opportunities of application. The combination of these novel ideas to improve the state-of-the-art of power systems optimization is an important direction of future work. On the other hand, low cost information technology (HPC and Big Data) as well as progresses in machine learning and randomized algorithms offer other enabling approaches to apply optimization techniques in power systems. We suggest that the research community should further focus on the proper formulation of power system optimization problems with the help of power system experts, and develop more intensively fruitful collaborations with researchers in applied mathematics and computer science to determine the most effective solution strategies for these problems. At the same time, we think that more systematic investments in a more effective use of modern information technologies, especially in the context of high-performance computing and massive data exploitation should be made by the power systems industry.

References

1. R.T. Rockafellar, R.J.-B. Wets. *Scenarios and policy aggregation inoptimization under uncertainty*. Mathematics of Operations Research,16(1):119–147, 1991.
2. F. C. Schweppe. *Power systems "2000": hierarchical control strategies*. Spectrum, IEEE, Volume:15, Issue: 7, July 1978.
3. *e-HighWay2050: Modular Development Plan of the Pan-European Transmission System 2050*. http://www.e-highway2050.eu/
4. *GARPUR: Generally Accepted Reliability Principle with Uncertainty modelling and through probabilistic Risk assessment*. http://www.garpur-project.eu/
5. A. Bemborad and M. Morari. *Control of systems integrating logic, dynamics, and constraints*. Automatica, 35:407–427, 1999.
6. L. Wehenkel, M. Pavella, E. Euxibie, B. Heilbronn. *Decision tree basedtransient stability method – a case study*. IEEE Trans. on Power Syst.,9(1):459–469, 1994.

3.22 A classification-based perspective to optimal policy design for a Markov Decision Process

Maria Prandini (Politecnico di Milano University, IT)

Joint work of Prandini, Maria; Manganini, Giorgio; Piroddi, Luigi; Pirotta, Matteo; Restelli, Matteo
Main reference M. Pirotta, G. Manganini, L. Piroddi, M. Prandini, M. Restelli, "A particle-based policy for the optimal control of Markov decision processes," in Proc. of the 19th IFAC World Congress 2014, Part 1, pp. 10518–10523, 2014.
URL http://dx.doi.org/10.3182/20140824-6-ZA-1003.01987

Classical approximate dynamic programming techniques based on state space gridding become computationally impracticable for high-dimensional problems. Policy search techniques cope with this curse of dimensionality issue by searching for the optimal control policy in a restricted parameterized policy space. In this work, we focus on the case of discrete action space and introduce a novel policy parameterization inspired by the nearest-neighbor classification method. Particles are adopted to describe the map from the state space to the action space, each particle representing a region of the state space that is mapped into a certain action. The locations and actions associated to the particles describing a policy are tuned through the policy gradient method with parameter-based exploration. The task of selecting an appropriately sized set of particles is solved through an iterative policy building scheme, that adds new particles to improve the policy performance and is also capable of removing redundant particles. Experiments demonstrate the scalability of the proposed approach as the dimensionality of the state space grows.

3.23 Time Unbounded Search for Error Trajectories of Complex Dynamical Systems

Stefan Ratschan (Academy of Science – Prague, CZ)

Joint work of Kuřátko, Jan; Ratschan, Stefan
Main reference J. Kuřátko, S. Ratschan, "Combined Global and Local Search for the Falsification of Hybrid Systems," in Proc. of the 12th Int'l Conf. on Formal Modeling and Analysis of Timed Systems (FORMATS'14), LNCS, Vol. 8711, pp. 146–160, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-10512-3_11

We present an algorithm that searches for an error trajectory of a given dynamical system. An error trajectory is a trajectory that starts in a given set of initial states and ends in a given set of states that is considered not to be safe. Unlike other methods, we do not require a user-provided upper bound on the length of this trajectory in terms of time.

Our method takes a black box model of the dynamical system that allows us to simulate the system from a given starting point for a certain amount of time, and that, if available, also returns the sensitivity of the computed trajectory w.r.t. this starting point.

3.24 Meshed Data- and Model- Driven Frameworks for Stochastic Hybrid Network Dynamics

Sandip Roy (Washington State University – Pullman, US)

Joint work of Roy, Sandip; Jackeline Abad Torres; Justin Valdez; Mengran Xue; Rahul Dhal; Yan Wan

Management of electric power networks depends on tracking and forecasting complex stochastic hybrid dynamics at several different scales. Mathematical models have been developed for many of these complex phenomena, but often cannot by themselves provide accurate predictions and forecasts. However, as cyber- capabilities are being integrated into the power grid, a wealth of data – ranging from synchrophasor measurements to smart meters and weather forecasts – is becoming available to stakeholders. This talk will explore how such data sources can be used in tandem with formal models to achieve situational awareness (monitoring and forecasting), using two concrete examples at different scales. First, we will briefly discuss how stochastic automata models can be used along with weather forecast data for wide-area renewable-generation forecasting. Second, we will present a technique for locating line faults in a power network, using models for transient dynamics together with synchrophasor data. A common theory for network estimation and control underlying these results will be briefly overviewed.

3.25 Model checking probabilistic hybrid automata

Jeremy Sproston (University of Turin, IT)

This talk gives an overview of the probabilistic hybrid automata, which are classical (non-deterministic) hybrid automata extended with probabilistic choice over the discrete part of the system in question. This formalism can be used for the modelling of, for example, faulty components or timed, randomised protocols. Model-checking analysis of probabilistic hybrid automata is considered. It is shown that results on obtaining finite-state models that are equivalent to hybrid automata can be lifted to the probabilistic hybrid automata case. Furthermore, a method in which probabilistic hybrid automata can be used as abstract models for more general classes of stochastic hybrid system is presented.

3.26 Operation of a transmission network as a hybrid system

Martin Střelec (UWB – Pilsen, CZ)

Joint work of Střelec, Martin; Cerny, Vaclav; Janecek, Petr

Transmission grids are large power networks interconnecting power generation with consumption and mainly constitute backbone power grids on national levels. A network operator (e. g. TSO) ensures safe and reliable transmission of electricity and its desired quality. Balance between power generation and demand is one of most important indicator in network safety assessment. For power balancing, network operators rely on ancillary services, which eliminate power differences between generated and consumed power. The proposed presentation focuses on hybrid modeling of transmission network operations and associated Monte Carlo simulation techniques. Operation of a transmission network is modelled by evolution of the power difference which is a process affected by continuous (e. g. load evolution, automatic generation control), discrete (e. g. forced generators outages, power reserves activations) and stochastic factors (e. g. load and renewable generation volatility). Therefore stochastic hybrid systems represents promising framework for modeling of transmission grid operation. Introduced hybrid model is utilized in a traditional Monte Carlo approach for reliability assessment of ancillary services. Some of ancillary services are contracted on long term basis (typically secondary control reserves). Reasonable volume determination of these ancillary services has significant economic impact on operators. Each network operator has their own methodology for long term contracting of ancillary services. Reliability assessment consists in verification whether contracted ancillary services are sufficient for reliable operation of transmission networks.

3.27 Correct-by-construction synthesis of software-based protocols for open, networked systems

Ufuk Topcu (University of Pennsylvania, US)

We discuss specification and synthesis of software-based control protocols that dynamically reconfigure a network as the environment in which the network operates changes. We consider power networks on aircraft as a driving application. In this case, the network needs to reconfigure in order to guarantee, for example, satisfactory delivery of electric power to loads as the flight conditions and the health status of the generators and other equipment change. We present samples of our recent work on correct-by-construction synthesis formulated as two-player, temporal logic-constrained games and stochastic, two- player games along with experiments on an academic-scale testbed.

3.28 Lazy Determinisation for Quantitative Model Checking

Lijun Zhang (Chinese Academy of Sciences, CN)

Joint work of Ernst Moritz Hahn; Guangyuan Li; Sven Schewe; Andrea Turrini; Zhang, Lijun
Main reference E. M. Hahn, G. Li, S. Schewe, L. Zhang, "Lazy Determinisation for Quantitative Model Checking," arXiv:1311.2928v1 [cs.LO], 2013.
URL http://arxiv.org/abs/1311.2928v1

The bottleneck in the quantitative analysis of Markov chains and Markov decision processes against specifications given in LTL or as some form of nondeterministic Buechi automata is the inclusion of a determinisation step of the automaton under consideration. In this paper we argue that applying this step in full generality is only required in some places, and can often be circumvented by subset and breakpoint constructions. We have implemented our approach – both explicit and symbolic versions – in a prototype tool. Our experiments show that our prototype can compete with mature tools like PRISM. Details can be found in our report [1].

References

1 Ernst Moritz Hahn and Guangyuan Li and Sven Schewe and Lijun Zhang. *Lazy Determinisation for Quantitative Model Checking*. CORR abs/1311.2928, December, 2003

3.29 Revisiting the mathematical modeling of power networks

Arjan van der Schaft (University of Groningen, NL)

Main reference S. Fiaz, D. Zonetti, R. Ortega, J. M. A. Scherpen, A. J. van der Schaft, "A port-Hamiltonian approach to power network modeling and analysis", European Journal of Control,19(6):477–485, 2013.
URL http://dx.doi.org/10.1016/j.ejcon.2013.09.002

Traditionally the power network engineering literature has refrained from developing 'complete' structure-preserving models of power networks. Instead one has emphasized phasor descriptions of the network, swing equations for the generators and/or the modeling of generators as ‘voltage sources behind a reactance’. There have been, and are, good reasons for doing this. On the other hand, with all the new questions posed in recent power network research developments, it seems promising to take a critical look at the modeling paradigms, in order to facilitate analysis and control. In this talk I will argue that a full time-domain modeling approach can be taken, which emphasizes the physical structure of the network and its components, and in particular makes ‘energy’ and ‘power flow’ explicit. Starting point of this approach will be a port-Hamiltonian formulation of the classical 8-dimensional model of the synchronous generator.

4 Working Groups

In the following we summarise the main points that have arisen from the breakout sessions and the final discussion. There have been three breakout sessions, focusing respectively on

- modelling issues in energy/power systems;
- simulation issues in energy/power systems;
- demand response: control and verification.

4.1 Modelling issues in energy/power systems

Alessandro Abate, Martin Střelec

The session has zoomed in on a number of topics, and raised related pressing issues.

Gap in models

- Study of steady-state (power flow), and of optimal power flow (OPF)
- Communications
- Interactions with the physical system
- Interactions in the infrastructures for gas and electricity

Uncertainty in models

- Presence of different forms of uncertainty: load amount and composition, load control availability, renewable generation, communication delays, random events (such as line and generator tripping), controllable resource in the case of demand response
- What are the right representations for power flow scheduling, analysis of its dynamics, its control?
- How can the effects of uncertainty on system dynamic performance be assessed (considering nonlinear, non-smooth behaviour)?

Timescale of interest in models

- Adaption of models for different time scales
- Phasor versus time-domain modelling for dynamic analysis
- Timescale differences can be particularly troublesome for micro-grid analysis

Load modelling

- Component models versus performance models, versus time-series models
- Distribution network models: aggregation of heterogeneous loads together with distributed generation
- Presence of uncertainty
- Models for control of individual loads, versus those for aggregated loads

Power electronic devices

- Presence of AC/DC, DC/AC conversion
- Very fast response
- Control-based rather than physics-based behaviour

4.2 Simulation issues in energy/power systems

Alessandro Abate, Martin Střelec

This group has focused on a number of topics, as discussed below.

Quality aspects of models

Model flexibility, namely the the ability to make user-defined changes to a model, is directly linked to the white-box vs. black-box dichotomy. Accuracy, namely the ability to show instabilities if the modelled system is unstable, is important, as much as computational complexity and portability are. The latter aspect in particular deals with the possible migration between simulation tools, platforms, and runtime environments.

Purpose-specific views on models

The internal structure and the model representation (stochastic, deterministic etc.) has an impact on the processes of 1. testing validation of models, and 2. optimisation of/within the model.

Determining the right models

This discussion focuses on a case-specific view on models. This can be based on the use cases or applications, where different model quality aspects become relevant, such as 1. planning vs. operation use cases; 2. design of control/protection systems; and 3. time-domain (non-linear) vs. frequency-domain aspects. In general, the more transient/dynamic the use case, the more transparent the model needs to be (from black- to grey- to white-box).

This theme has covered a discussion on the inclusion of noise into the models of power systems (in order to trigger certain effects that would otherwise be undetectable).

Necessary tools

Starting from a discussion on how the engineering and the integration of models is quite expensive, the following aspects have been touched upon:

1. automatic reduction/abstraction/simplification of models;
2. merging/fusion of (white-box) models;
3. co-simulation; and
4. validation (benchmarking against reference models).

Open questions

The following issues remain:

1. how do you validate or benchmark a system, without the data or the knowledge of the "real" underlying system (e. g. we are making highly detailed models of distribution systems that are currently not measured at all)?
2. can you derive model/application-specific characteristics or metrics?
3. can you tune models online as the data is gathered? Model-specific processes and tools are needed!

4.3 Demand response: control and verification

Alessandro Abate, Martin Střelec

Briefly stated, demand response has to do with the reaction within an energy-consuming building or by a consumer to requests to adjust consumption. The request can be direct (e. g. direct on/off control of HVAC system) or indirect (through pricing mechanisms).

Broadly, topics discussed have been the following:

- whether optimality issues can be properly addressed with optimal control or MPC techniques;
- how high-gain and delay in feedback loop in response to the pricing signal could result in instability;
- issues of privacy, due to load profiles being transmitted, particularly within bidirectional collaborative energy management;
- possible trade-offs between performance degradation and responsiveness to pricing stimuli;
- issues of faults diagnosis and of optimal control/scheduling for maintenance in HVAC systems;
- issues and opportunities in risk hedging/diversification of renewable sources;
- issues of predictability of renewables (in a short time scale), to be potentially matched by demand response?
- timely acquisition, processing and decision making based on "real-time" info.

Additional points raised in the discussions have touched upon the following themes:

- intelligent management/optimization and storage at local/global level;
- management of mobile assets (such as meters, devices, cars), with their user-friendliness;
- flexible policy for data management, dealing with security, privacy, and trust;
- the importance of interoperability and of open cross-layer collaborative approaches;
- issues of reliability and of scalability;

- the required support for and development of qualified open standards;
- the requirement of large-scale simulation, modelling, and risk analysis tools;
- the importance of business analytics (prediction, visualization, etc.);
- and finally the need for real-world trials and experiences.

Participants

- Alessandro Abate
University of Oxford, GB
- Erika Ábrahám
RWTH Aachen, DE
- Christel Baier
TU Dresden, DE
- Calin A. Belta
Boston University, US
- Marc Bouissou
Ecole Centrale Paris, FR
- Ana Busic
INRIA – Paris, FR
- Claudio De Persis
University of Groningen, NL
- Florian Dörfler
ETH Zürich, CH
- Martin Fränzle
Universität Oldenburg, DE
- Sicun Gao
Carnegie Mellon University, US
- Boudewijn Haverkort
University of Twente, NL
- Holger Hermanns
Universität des Saarlandes, DE
- Ian Hiskens
University of Michigan – Ann Arbor, US
- Ondřej Holub
Honeywell Prague Lab., CZ
- Gabriela Hug-Glanzmann
Carnegie Mellon University, US
- George B. Huitema
University of Groningen, NL
- Anak Agung Julius
Rensselaer Polytechnic Inst., US
- Maryam Kamgarpour
ETH Zürich, CH
- Stamatis Karnouskos
SAP Research – Karlsruhe, DE
- Joost-Pieter Katoen
RWTH Aachen, DE
- David Kleinhans
NEXT ENERGY – Oldenburg, DE
- Jan Krčál
Universität des Saarlandes, DE
- Rom Langerak
University of Twente, NL
- Mats Larsson
ABB Corporate Research – Baden-Dättwil, CH
- Sebastian Lehnhoff
OFFIS – Oldenburg, DE
- Bernard Lesieutre
University of Wisconsin – Madison, US
- Roland P. Malhame
Polytechnique Montreal, CA
- Kostas Margellos
University of California – Berkeley, US
- Johanna Mathieu
University of Michigan – Ann Arbor, US
- Patrick Panciatici
RTE Energy – France, FR
- Maria Prandini
Politecnico di Milano Univ., IT
- Stefan Ratschan
Academy of Science – Prague, CZ
- Anne Remke
University of Twente, NL
- Sandip Roy
Washington State University – Pullman, US
- Jeremy Sproston
University of Turin, IT
- Martin Střelec
UWB – Pilsen, CZ
- Ufuk Topcu
University of Pennsylvania, US
- Arjan van der Schaft
University of Groningen, NL
- Lijun Zhang
Chinese Academy of Sciences, CN

Report from Dagstuhl Seminar 14442

Symbolic Execution and Constraint Solving

Edited by

Cristian Cadar[1], Vijay Ganesh[2], Raimondas Sasnauskas[3], and Koushik Sen[4]

1 **Imperial College London, GB,** c.cadar@imperial.ac.uk
2 **University of Waterloo, CA,** vganesh@uwaterloo.ca
3 **University of Utah, US,** rsas@cs.utah.edu
4 **University of California, Berkeley, US,** ksen@cs.berkeley.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 14442 "Symbolic Execution and Constraint Solving", whose main goals were to bring together leading researchers in the fields of symbolic execution and constraint solving, foster greater communication between these two communities and exchange ideas about new research directions in these fields.

There has been a veritable revolution over the last decade in the symbiotic fields of constraint solving and symbolic execution. Even though key ideas behind symbolic execution were introduced more than three decades ago, it was only recently that these techniques became practical as a result of significant advances in constraint satisfiability and scalable combinations of concrete and symbolic execution. Thanks to these advances, testing and analysis techniques based on symbolic execution are having a major impact on many sub-fields of software engineering, computer systems, security, and others. New applications such as program and document repair are being enabled, while older applications such as model checking are being super-charged. Additionally, significant and fast-paced advances are being made in research at the intersection of traditional program analysis, symbolic execution and constraint solving. Therefore, this seminar brought together researchers in these varied fields in order to further facilitate collaborations that take advantage of this unique and fruitful confluence of ideas from the fields of symbolic execution and constraint solving.

Seminar October 27–30, 2014 – http://www.dagstuhl.de/14442

1998 ACM Subject Classification D.2.4 Software/Program Verification (Formal methods), D.2.5 Testing and Debugging (Symbolic execution, Testing tools (e.g., data generators, coverage testing), Tracing), B.2.3 Reliability, Testing, and Fault-Tolerance (Test generation)

Keywords and phrases Symbolic Execution, Software Testing, Automated Program Analysis, Constraint Solvers

Digital Object Identifier 10.4230/DagRep.4.10.98

1 Executive Summary

Cristian Cadar
Vijay Ganesh
Raimondas Sasnauskas
Koushik Sen

Symbolic execution has garnered a lot of attention in recent years as an effective technique for generating high-coverage test suites, finding deep errors in complex software applications, and more generally as one of the few techniques that is useful across the board in myriad software engineering applications. While key ideas behind symbolic execution were introduced more than three decades ago, it was only recently that these techniques became practical as a result of significant advances in constraint satisfiability and scalable combinations of concrete and symbolic execution. The result has been an explosion in applications of symbolic execution techniques in software engineering, security, formal methods and systems research. Furthermore, researchers are combining symbolic execution with traditional program analysis techniques in novel ways to address longstanding software engineering problems. This in turn has led to rapid developments in both constraint solvers and symbolic execution techniques, necessitating an in-depth exchange of ideas between researchers working on solvers and symbolic techniques, best accomplished through dedicated workshops.

Hence, one of the main goals of this Dagstuhl seminar was to bring together leading researchers in the fields of symbolic execution and constraint solving, foster greater communication between these two communities and discuss new research directions in these fields. The seminar had 34 participants from Canada, France, Germany, Norway, Singapore, South Africa, Spain, Switzerland, The Netherlands, United Kingdom and United States, from both academia, research laboratories, and the industry. More importantly, the participants represented several different communities, with the topics of the talks and discussions reflecting these diverse interests: testing, verification, security, floating point constraint solving, hybrid string-numeric constraints, debugging and repair, education, and commercialization, among many others.

2 Table of Contents

3 Overview of Talks

3.1 Automated White-Box Testing Beyond Branch Coverage

Sébastien Bardin (CEA LIST, FR)

Joint work of Bardin, Sébastien; Kosmatov, Nikolai; Delahaye, Mickaël; Chebaro, Omar
Main reference S. Bardin, N. Kosmatov, F. Cheynier, "Efficient Leveraging of Symbolic Execution to Advanced Coverage Criteria," in Proc. of the 2014 IEEE 7th Int'l Conf. on Software Testing, Verification and Validation (ICST'14), pp. 173–182, IEEE, 2014.
URL http://dx.doi.org/10.1109/ICST.2014.30

Automated white-box testing is a major issue in software engineering. The last decade has seen tremendous progress in Automatic test data generation thanks to Symbolic Execution techniques. Yet, these promising results are currently limited to very basic coverage criteria such as statement coverage or branch coverage, while many more criteria can be found in the literature. Moreover, other important issues in structural testing, such as infeasible test requirement detection, have not made so much progress.

In order to tackle these problems, we rely on a simple and concise specification mechanism for (structural) coverage criteria, called labels. Labels are appealing since they can faithfully encode many existing coverage criteria, allowing to handle all of them in a unified way. They are the corner stone of FRAMA-C/LTEST, a generic and integrated toolkit for automated white-box testing of C programs, providing automatic test generation and detection of infeasible test requirements through a combination of static and dynamic analyzes.

We will overview the platform and focus our presentation on the following points: (1) present a new and efficient Symbolic Execution algorithm dedicated to label coverage, and (2) describe recent results on the sound detection of infeasible test requirements.

References

1 Sébastien Bardin, Nikolai Kosmatov, François Cheynier: Efficient Leveraging of Symbolic Execution to Advanced Coverage Criteria. In: ICST 2014. IEEE (Los alamitos)
2 Sébastien Bardin, Omar Chebaro, Mickaël Delahaye, Nikolai Kosmatov: An All-in-One Toolkit for Automated White-Box Testing. In: TAP 2014. Springer (Heidelberg)

3.2 Dynamic Symbolic Execution: State of the Art, Applications and Challenges

Cristian Cadar (Imperial College London, GB)

In this talk, I start by presenting the state of the art in dynamic symbolic execution, including its main enablers, namely mixed concrete/symbolic execution, better and faster constraint solvers and novel path exploration algorithms. I then survey several applications of symbolic execution, including bug finding, security, high-coverage test generation, software debugging, patch and document recovery, etc. I finish the talk by discussing some of the ongoing challenges in terms of path explosion, verification, concurrency and constraint solving.

3.3 Reaching Verification with Systematic Testing

Maria Christakis (ETH Zürich, CH)

Joint work of Christakis, Maria; Godefroid, Patrice
Main reference M. Christakis, P. Godefroid, "Proving memory safety of the ANI Windows image parser using compositional exhaustive testing," in Proc. of the 16th Int'l Conf. on Verification, Model Checking, and Abstract Interpretation (VMCAI'15), LNCS, Vol. 8931, pp. 370–389, Springer, 2015.
URL http://dx.doi.org/10.1007/978-3-662-46081-8_21

We describe how we proved memory safety of a complex Windows image parser written in low-level C in only three months of work and using only three core techniques, namely (1) symbolic execution at the x86 binary level, (2) exhaustive program path enumeration and testing, and (3) user-guided program decomposition and summarization. As a result of this work, we are able to prove, for the first time, that a Windows image parser is memory safe, that is, free of any buffer-overflow security vulnerabilities, modulo the soundness of our tools and several additional assumptions regarding bounding input-dependent loops, fixing a few buffer-overflow bugs, and excluding some code parts that are not memory safe by design.

3.4 Superoptimizing LLVM

Peter Collingbourne (Google Inc. – Mountain View, US)

Joint work of Collingbourne, Peter; Regehr, John; Sasnauskas, Raimondas; Taneja, Jubi; Chen, Yang; Ketema, Jeroen

Compiler development, as with any engineering task, involves certain tradeoffs. In particular, the tradeoff between the expected performance improvement of a specific optimization and the engineer time required to implement it can be difficult to assess, especially given the long tail of potential uncaught optimizations that may be lurking within a large codebase.

Souper is an open source superoptimizer based on LLVM which can automatically extract potential peephole optimizations from real C and C++ programs, use a SMT solver to verify their correctness, apply them automatically, and use static and dynamic profiling to identify the most profitable optimizations. It pushes the state of the art in optimizer development forward in two primary ways: directly, by instantly providing a way for users to automatically apply a variety of previously unimplemented peephole optimizations to their code; and indirectly, by allowing compiler developers to focus their energy on implementing the most profitable optimizations in the baseline compiler.

This talk described Souper, giving details of the translation from LLVM to SMT predicates, and results from compiling LLVM with Souper.

3.5 Experiences with SMT in the GPUVerify Project

Alastair F. Donaldson (Imperial College London, GB)

Joint work of Bardsley, Ethel; Betts, Adam; Chong, Nathan; Collingbourne, Peter; Deligiannis, Pantazis; Donaldson, Alastair F.; Ketema, Jeroen; Liew, Daniel; Qadeer, Shaz
Main reference E. Bardsley, A. Betts, N. Chong, P. Collingbourne, P. Deligiannis, A. F. Donaldson, J. Ketema, D. Liew, S. Qadeer, "Engineering a Static Verification Tool for GPU Kernels," in Proc. of the 26th Int'l Conf. on Computer Aided Verification (CAV'14), LNCS, Vol. 8559, pp. 226–242, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-08867-9_15

The GPUVerify project has investigated techniques for automatically proving freedom from data races in GPU kernels, implemented in a practical tool. To achieve efficiency and a relatively high degree of automation, we have put a lot of effort into designing encodings of properties of GPU kernels into SMT formulas that avoid the use of quantifiers, which are notoriously hard to reason about automatically. In our experiments with various encodings, using the Z3 and CVC4 solvers, we have observed a great deal of variation in response time between solvers, and across encodings with respect to a single solver. In this talk I gave an overview of one SMT encoding of the data race-freedom property for GPU kernels, and presented experimental results illustrating the variation in results we have observed. The results demonstrate the need to evaluate optimizations to a verification technique with respect to large set of benchmarks, and to evaluate SMT-based optimizations using multiple solvers.

3.6 On the Challenges of Combining Search-Based Software Testing and Symbolic Execution

Juan Pablo Galeotti (Universität des Saarlandes, DE)

Joint work of Galeotti, Juan Pablo; Gordon Fraser; Andrea Arcuri; Matthias Höschele; Andreas Zeller
Main reference J. P. Galeotti, G. Fraser, A. Arcuri, "Improving search-based test suite generation with dynamic symbolic execution," in Proc. of the 2013 IEEE 24th Int'l Symp. on Software Reliability Engineering (ISSRE'13), pp. 360–360, IEEE, 2013.
URL http://dx.doi.org/10.1109/ISSRE.2013.6698889

In recent years, there has been a tremendous advance in Search-Based Software Testing and Symbolic Execution. Search-based testing (SBST) can automatically generate unit test suites for object oriented code, but may struggle to generate specific values necessary to cover difficult parts of the code. Symbolic execution (SE) efficiently generates such specific values, but may struggle with complex datatypes, in particular those that require sequences of calls for construction.

In this talk I will present a hybrid approach for automatic unit-test generation that integrates the best of both worlds [1]. Also, I will focus on the challenges we recently faced when applying SBST and SE to multi-layered software comprised of an object-oriented upper layer and a native-platform dependent native layer [2].

References

1 Juan Pablo Galeotti, Gordon Fraser, Andrea Arcuri: *Improving search-based test suite generation with dynamic symbolic execution*. In ISSRE 2013, pages 360–369
2 Matthias Höschele, Juan Pablo Galeotti, Andreas Zeller: *Test generation across multiple layers*. In SBST 2014, pages 1–4

3.7 Impact of Community Structure on SAT Solver Performance

Vijay Ganesh (University of Waterloo, CA)

Joint work of Newsham, Zack; Ganesh, Vijay; Fischmeister, Sebastian; Audemard, Gilles; Simon, Laurent
Main reference Zack Newsham, V. Ganesh, S. Fischmeister, G. Audemard, L. Simon, "Impact of Community Structure on SAT Solver Performance," in Proc. of the 17th Int'l Conf. on Theory and Applications of Satisfiability Testing (SAT'14), LNCS, Vol. 8561, pp. 252–268, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-09284-3_20

Modern CDCL SAT solvers routinely solve very large industrial SAT instances in relatively short periods of time. It is clear that these solvers somehow exploit the structure of real-world instances. However, to-date there have been few results that precisely characterise this structure. In this paper, we provide evidence that the community structure of real-world SAT instances is correlated with the running time of CDCL SAT solvers. It has been known for some time that real-world SAT instances, viewed as graphs, have natural communities in them. A community is a sub-graph of the graph of a SAT instance, such that this sub-graph has more internal edges than outgoing to the rest of the graph. The community structure of a graph is often characterised by a quality metric called Q. Intuitively, a graph with high-quality community structure (high Q) is easily separable into smaller communities, while the one with low Q is not. We provide three results based on empirical data which show that community structure of real-world industrial instances is a better predictor of the running time of CDCL solvers than other commonly considered factors such as variables and clauses. First, we show that there is a strong correlation between the Q value and Literal Block Distance metric of quality of conflict clauses used in clause-deletion policies in Glucose-like solvers. Second, using regression analysis, we show that the the number of communities and the Q value of the graph of real-world SAT instances is more predictive of the running time of CDCL solvers than traditional metrics like number of variables or clauses. Finally, we show that randomly-generated SAT instances with $0.05 \leq Q \leq 0.13$ are dramatically harder to solve for CDCL solvers than otherwise.

3.8 Combining FSM Modeling and Bit-Vector Theories to Solve Hybrid String-Numeric Constraints

Indradeep Ghosh (Fujitsu Labs of America Inc. – Sunnyvale, US)

Joint work of Ghosh, Indradeep; Li, Goudong
Main reference G.Li, I. Ghosh, "PASS: String Solving with Parameterized Array and Interval Automaton," in Proc. of the 9th Int'l Haifa Verification Conference (HVC'13), LNCS, Vol. 8244, pp. 15–31, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-319-03077-7_2

This talk focused on the challenges and techniques for solving hybrid String-Numeric constraints. These types of constraints arise as path conditions during symbolic execution of industrial application, especially enterprise applications written in Java and JavaScript programming languages. Two types of solving techniques were presented: Finite state machine modeling of strings and Bit-vector models of strings. Their pros and cons were discussed and ways to merge them using parameterized arrays and interval automaton were presented.

3.9 Symbolic Path-Oriented Test Data Generation for Floating-Point Programs

Arnaud Gotlieb (Simula Research Laboratory – Lysaker, NO)

Joint work of Bagnara, Roberto; Carlier, Matthieu; Gori, Roberta; Gotlieb, Arnaud
Main reference R. Bagnara, M. Carlier, R. Gori, A. Gotlieb, "Symbolic path-oriented test data generation for floating-point programs," in Proc. of the 2013 IEEE 6th Int'l Conf. on Software Testing, Verification and Validation (ICST'13), pp. 1–10, IEEE, 2013.
URL http://dx.doi.org/10.1109/ICST.2013.17

Verifying critical numerical software involves the generation of test data for floating-point intensive programs. As the symbolic execution of floating-point computations presents significant difficulties, existing approaches usually resort to random or search-based test data generation. However, without symbolic reasoning, it is almost impossible to generate test inputs that execute many paths with floating-point computations. Moreover, constraint solvers over the reals or the rationals do not handle the rounding errors. In this paper, we present a new version of FPSE, a symbolic evaluator for C program paths, that specifically addresses this problem. The tool solves path conditions containing floating-point computations by using correct and precise projection functions. This version of the tool exploits an essential filtering property based on the representation of floating-point numbers that makes it suitable to generate path-oriented test inputs for complex paths characterized by floating-point intensive computations. The paper reviews the key implementation choices in FPSE and the labeling search heuristics we selected to maximize the benefits of enhanced filtering. Our experimental results show that FPSE can generate correct test inputs for selected paths containing several hundreds of iterations and thousands of executable floating-point statements on a standard machine: this is currently outside the scope of any other symbolic execution test data generator tool.

References

1 R. Bagnara, M. Carlier, R. Gori, and A. Gotlieb. Symbolic path-oriented test data generation for floating-point programs. In Proc. of the 6th IEEE Int. Conf. on Software Testing, Verification and Validation (ICST'13), Luxembourg, Mar. 2013.
2 B. Botella, A. Gotlieb, and C. Michel. Symbolic execution of floating-point computations. Software Testing, Verification and Reliability (STVR), 16(2):97–121, June 2006

3.10 Segmented Symbolic Analysis

Wei Le (Iowa State University – Ames, US)

Main reference W. Le, "Segmented Symbolic Analysis," in Proc. of the 2013 35th Int'l Conf. on Software Engineering (ICSE '13), pp. 212–221, IEEE, 2013.
URL http://dx.doi.org/10.1109/ICSE.2013.6606567

Symbolic analysis is indispensable for software tools that require program semantic information at compile time. However, determining symbolic values for program variables related to loops and library calls is challenging, as the computation and data related to loops can have statically unknown bounds, and the library sources are typically not available at compile time. In this talk, I present segmented symbolic analysis, a hybrid technique that enables fully automatic symbolic analysis even for the traditionally challenging code of library calls and loops. The novelties of this work are threefold: 1) we flexibly weave symbolic and concrete executions on the selected parts of the program based on demand; 2) dynamic executions are performed on the unit tests constructed from the code segments to infer program semantics needed by static analysis; and 3) the dynamic information from multiple runs is aggregated via regression analysis. We developed the Helium framework, consisting of a static component that performs symbolic analysis and partitions a program, a dynamic analysis that synthesizes unit tests and automatically infers symbolic values for program variables, and a protocol that enables static and dynamic analyses to be run interactively and concurrently. Our experimental results show that by handling loops and library calls that a traditional symbolic analysis cannot process, segmented symbolic analysis detects 5 times more buffer overflows. The technique is scalable for real-world programs such as putty, tightvnc and snort.

3.11 Commercial Symbolic Execution

Paul Marinescu (Imperial College London, GB)

Symbolic execution is a powerful automatic testing technique which recently received a lot of attention in academic circles, but has not found its way into mainstream software engineering because reaching a balance between usability, effectiveness and resource requirements has proved elusive thus far. This talk looks at the challenges that need to be overcome to make symbolic execution a commercial success, comparing it with static analysis-based commercial solutions.

3.12 Solving Non-linear Integer Constraints Arising from Program Analysis

Albert Oliveras (Polytechnic University of Catalonia, ES)

Joint work of Larraz, Daniel; Oliveras, Albert; Rodriguez-Carbonell, Enric; Rubio, Albert
Main reference D. Larraz, A. Oliveras, E. Rodríguez-Carbonell, A. Rubio,“Minimal-Model-Guided Approaches to Solving Polynomial Constraints and Extensions,” in Proc. of the 17th Int'l Conf. on Theory and Applications of Satisfiability Testing (SAT'14), LNCS, Vol. 8561, pp. 333-350, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-09284-3_25

We present new methods for deciding the satisfiability of formulas involving integer polynomial constraints. In previous work we proposed to solve SMT(NIA) problems by reducing them to SMT(LIA): non-linear monomials are linearized by abstracting them with fresh variables and by performing case splitting on integer variables with finite domain. When variables do not have finite domains, artificial ones can be introduced by imposing a lower and an upper bound, and made iteratively larger until a solution is found (or the procedure times out). For the approach to be practical, unsatisfiable cores are used to guide which domains have to be relaxed (i.e., enlarged) from one iteration to the following one. However, it is not clear then how large they have to be made, which is critical.

Here we propose to guide the domain relaxation step by analyzing minimal models produced by the SMT(LIA) solver. Namely, we consider two different cost functions: the number of violated artificial domain bounds, and the distance with respect to the artificial domains. We compare these approaches with other techniques on benchmarks coming from constraint-based program analysis and show the potential of the method. Finally, we describe how one of these minimal-model-guided techniques can be smoothly adapted to deal with the extension Max-SMT of SMT(NIA) and then applied to program termination proving.

3.13 Constraint Solving in Symbolic Execution

Hristina Palikareva (Imperial College London, GB)

Joint work of Palikareva, Hristina; Cadar, Cristian
Main reference H. Palikareva, C. Cadar, "Multi-solver Support in Symbolic Execution," in Proc. of the 25th Int'l Conf. on Computer Aided Verification (CAV'13), LNCS, Vol. 8044, pp. 53–68, Springer, 2013; pre-print available from author's webpage.
URL http://dx.doi.org/10.1007/978-3-642-39799-8_3
URL http://srg.doc.ic.ac.uk/files/papers/klee-multisolver-cav-13.pdf

Dynamic symbolic execution is an automated program analysis technique that employs an SMT solver to systematically explore paths through a program. It has been acknowledged in recent years as a highly effective technique for generating high-coverage test suites as well as for uncovering deep corner-case bugs in complex real-world software, and one of the key factors responsible for that success are the tremendous advances in SMT-solving technology. Nevertheless, constraint solving remains one of the fundamental challenges of symbolic execution, and for many programs it is the main performance bottleneck.

In this talk, we will present the results reported in our CAV 2013 paper on integrating support for multiple SMT solvers in the dynamic symbolic execution engine KLEE. In particular, we will outline the key characteristics of the SMT queries generated during symbolic execution, describe several high-level domain-specific optimisations that KLEE employs to exploit those characteristics, introduce an extension of KLEE that uses a number of state- of-the-art SMT solvers (Boolector, STP and Z3) through the metaSMT solver framework, and compare the solvers' performance when run on large sets of QF_ABV queries obtained during the symbolic execution of real-world software. In addition, we will discuss several options for designing a parallel portfolio solver for symbolic execution tools. We will conclude the talk by proposing the introduction of a separate division at the annual SMT competition targeted specifically at symbolic execution tools.

3.14 Provably Correct Peephole Optimizations with Alive

John Regehr (University of Utah, US)

Compilers should not miscompile. Our work addresses problems in developing peephole optimizations that perform local rewriting to improve the efficiency of LLVM code. These optimizations are individually difficult to get right, particularly in the presence of undefined behavior; taken together they represent a persistent source of bugs. This paper presents Alive, a domain-specific language for writing optimizations and for automatically either proving them correct or else generating counterexamples. Furthermore, Alive can be automatically translated into C++ code that is suitable for inclusion in an LLVM optimization pass. Alive is based on an attempt to balance usability and formal methods; for example, it captures–but largely hides–the detailed semantics of three different kinds of undefined behavior in LLVM. We have translated more than 300 LLVM optimizations into Alive and, in the process, found that eight of them were wrong.

3.15 Symbolic Techniques for Program Debugging

Abhik Roychoudhury (National University of Singapore, SG)

Joint work of Roychoudhury, Abhik; Chandra, Satish
URL http://www.slideshare.net/roychoudhury/abhik-satishdagstuhl-40988809

In recent years, there have been significant advances in symbolic execution technology, driven by the increasing maturity of SMT and SAT solvers as well as by the availability of cheap compute resources. This technology has had a significant impact in the area of automatically finding bugs in software. In this tutorial, we review ways in which symbolic execution can be used not just for finding bugs in programs, but also in debugging them! In current practice, once a failure-inducing input has been found, humans have to spend a great deal of effort in determining the root cause of the bug. The reason the task is complicated is that a person has to figure out manually how the execution of the program on the failure-inducing input deviated from the "intended" execution of the program. We show that symbolic analysis can be used to help the human in this task in a variety of ways. In particular, symbolic execution helps to glean the intended program behavior via analysis of the buggy trace, analysis of other traces or other program versions. Concretely, the tutorial provides a background in symbolic execution, and then covers material from a series of recent papers (including papers by the authors) on determination of root cause of errors using symbolic techniques.

3.16 Generating Heap Summaries from Symbolic Execution

Neha Rungta (NASA – Moffett Field, US)

Joint work of Hillery, Ben; Mercer, Eric; Rungta, Neha; Person, Suzette

A fundamental challenge of using symbolic execution for software analysis has been the treatment of dynamically allocated data. State-of-the-art techniques have addressed this challenge through either (a) use of summaries that over-approximate possible heaps or (b) by materializing a concrete heap lazily during generalized symbolic execution. In this work, we present a novel heap initialization and analysis technique which takes inspiration from both approaches and constructs precise heap summaries lazily during symbolic execution. Our approach is 1) scalable: it reduces the points of non-determinism compared to generalized symbolic execution and explores each control-flow path only once for any given set of isomorphic heaps, 2) precise: at any given point during symbolic execution, the symbolic heap represents the exact set of feasible concrete heap structures for the program under analysis, and 3) expressive: the symbolic heap can represent recursive data structures. We demonstrate the precision and scalability of our approach by implementing it as an extension to the Symbolic PathFinder framework for analyzing Java programs.

3.17 MultiSE: Multi-Path Symbolic Execution using Value Summaries

Koushik Sen (University of California, Berkeley)

Joint work of Sen, Koushik; Necula, George; Gong, Liang; Choi, Wontae
Main reference K. Sen, G. Necula, L. Gong, P. W. Choi, "MultiSE: Multi-Path Symbolic Execution using Value Summaries," Technical Report, UCB/EECS-2014-173, University of California, Berkeley, 2014.
URL http://www.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-173.html

Concolic testing of dynamic symbolic execution (DSE) has been proposed recently to effectively generate test inputs for real-world programs. Unfortunately, concolic testing techniques do not scale well for large realistic programs, because often the number of feasible execution paths of a program increases exponentially with the increase in the length of an execution path.

In this talk, I will describe MultiSE, a new technique for merging states incrementally during symbolic execution, without using auxiliary variables. The key idea of MultiSE is based on an alternative representation of the state, where we map each variable, including the program counter, to a set of guarded symbolic expressions called a value summary. MultiSE has several advantages over conventional DSE and state merging techniques: 1) value summaries enable sharing of symbolic expressions and path constraints along multiple paths, 2) value-summaries avoid redundant execution, 3) MultiSE does not introduce auxiliary symbolic values, which enables it to make progress even when merging values not supported by the constraint solver, such as floating point or function values.

We have implemented MultiSE for JavaScript programs in a publicly available open-source tool. Our evaluation of MultiSE on several programs shows that MultiSE can run significantly faster than traditional symbolic execution.

3.18 The Symbiosis of Network Testing and Symbolic Execution

Oscar Soria Dustmann (RWTH Aachen University, DE)

Creating new adaptive Internet technologies as envisioned by the MAKI project requires the interaction of a plethora of different modules, submodules and hardware platforms. Implementation-defined behaviour and the inherent concurrency and communication delay of these systems are the primary sources of many potential types of errors. As can already be observed in current highly distributed systems such errors tend to hide in sometimes quite obscure corner-cases. For example, a catastrophic protocol lock-up might manifest only for an unlikely, unexpected packet reordering. In addition, incompatibilities stemming from the heterogeneity of subsystems, demand further attention for testing approaches. This work aims at devising methodologies that address particularly timing and heterogeneity related issues in distributed and networked systems, with a focus on event-driven software.

3.19 Symbolic Execution and Model Counting

Willem Visser (Stellenbosch University – Matieland, ZA)

Joint work of Visser, Willem; Filieri, Antonio; Pasareanu, Corina; Dwyer, Matt; Geldenhuys, Jaco
Main reference A. Filieri, C. S. Păsăreanu, W. Visser, J. Geldenhuys, "Statistical symbolic execution with informed sampling," in Proc. of the 22nd ACM SIGSOFT Int'l Symp. on Foundations of Software Engineering (FSE'14), pp. 437–448, ACM, 2014.
URL http://dx.doi.org/10.1145/2635868.2635899

Symbolic execution has become a very popular means for analysing program behaviour. Traditionally it only considers whether paths are feasible or not. We argue there is a wealth of interesting new research directions that open up when we also consider how likely it is that a path is feasible. We show that by using the notion of model counting (how many solutions there are rather than just whether there is a solution to a constraint) we can calculate how likely an execution path through the code is to be executed. We show how this can be used to augment program understanding, calculate the reliability of the code and also as a basis for test coverage calculations. In order to speed up calculations we use the Green framework and will give a quick primer on how to use Green. Green is a framework to allow one to reuse results across analysis runs. It speeds up satisfiability checking and model counting at the moment, but can do anything you like in a flexible framework.

3.20 Feedback-Driven Dynamic Invariant Discovery

Lingming Zhang (University of Texas at Dallas, US)

Joint work of Zhang, Lingming; Yang, Guowei; Rungta, Neha; Person, Suzette; Khurshid, Sarfraz
Main reference L. Zhang, G. Yang, N. Rungta, S. Person, S. Khurshid, "Feedback-driven dynamic invariant discovery," in Proc. of the 2014 International Symposium on Software Testing and Analysis (ISSTA'14), pp. 362–372, ACM, 2014.
URL http://dx.doi.org/10.1145/2610384.2610389

Program invariants can help software developers identify program properties that must be preserved as the software evolves, however, formulating correct invariants can be challenging. In this work, we introduce iDiscovery, a technique which leverages symbolic execution to improve the quality of dynamically discovered invariants computed by Daikon. Candidate invariants generated by Daikon are synthesized into assertions and instrumented onto the program. The instrumented code is executed symbolically to generate new test cases that are fed back to Daikon to help further refine the set of candidate invariants. This feedback loop is executed until a fix-point is reached. To mitigate the cost of symbolic execution, we present optimizations to prune the symbolic state space and to reduce the complexity of the generated path conditions. We also leverage recent advances in constraint solution reuse techniques to avoid computing results for the same constraints across iterations. Experimental results show that iDiscovery converges to a set of higher quality invariants compared to the initial set of candidate invariants in a small number of iterations.

4 Thought-provoking Talks

In addition to the regular 25-minutes conference-style presentations, we encouraged the attendees to give 5-minutes talks on thought-provoking ideas. A key goal behind these short talks was to promote discussions among the attendees and to rethink our future research directions on symbolic execution and constraint solving. We got warm participation from the speakers and the audience. The topics discussed in these short talks include

- the application challenges of symbolic execution,
- the challenges in processing the test cases during symbolic execution of networked code,
- the idea of bringing static analysis and symbolic execution techniques together, and
- the application of symbolic execution in security.

Overall, we found these short-talk sessions to be engaging and fun. We encourage future workshop organizers to include similar sessions in their meetings.

5 Programme

Programme for Tuesday 28th of October

- Organizers: Welcome and introductions
- Cristian Cadar: Dynamic Symbolic Execution: State of the Art, Applications and Challenges
- Willem Visser: Symbolic Execution and Model Counting
- Koushik Sen: MultiSE: Multi-Path Symbolic Execution using Value Summaries
- Wei Le: Segmented Symbolic Analysis
- Peter Collingbourne: Superoptimizing LLVM
- John Regehr: Provably Correct Peephole Optimizations with Alive
- Arnaud Gotlieb: Symbolic Path-Oriented Test Data Generation for Floating-Point Programs
- Thought-provoking talks & discussions
- Indradeep Ghosh: Combining FSM Modeling and Bit-Vector Theories to Solve Hybrid String-Numeric Constraints
- Thought-provoking talks & discussions

Programme for Wednesday 29th of October

- Abhik Roychoudhury and Satish Chandra: Symbolic Techniques for Program Debugging
- Thought-provoking talks & discussions
- Vijay Ganesh: Impact of Community Structure on SAT Solver Performance
- Albert Oliveras: Solving Non-linear Integer Constraints Arising from Program Analysis
- Alastair F. Donaldson: Experiences with SMT in the GPUVerify Project
- Hristina Palikareva: Constraint Solving in Symbolic Execution
- Juan Pablo Galeotti: On the Challenges of Combining Search-Based Software Testing and Symbolic Execution
- Paul Marinescu: Commercial Symbolic Execution
- Maria Christakis: Reaching Verification with Systematic Testing
- Thought-provoking talks & discussions

Programme for Thursday 30th of October

- Neha Rungta: Generating Heap Summaries from Symbolic Execution
- Oscar Soria Dustmann: The Symbiosis of Network Testing and Symbolic Execution
- Lingming Zhang: Feedback-Driven Dynamic Invariant Discovery
- Istvan Haller: Symbolic execution in the field of security
- Sébastien Bardin: Automated White-Box Testing Beyond Branch Coverage
- Nicky Williams: Introduction to PathCrawler

Participants

- Sébastien Bardin
CEA LIST – Paris, FR
- Earl Barr
University College London, UK
- Cristian Cadar
Imperial College London, UK
- Satish Chandra
Samsung Electronics – San Jose, US
- Maria Christakis
ETH Zürich, CH
- Peter Collingbourne
Google Inc. – Mountain View, US
- Jorge R. Cuéllar
Siemens AG – München, DE
- Morgan Deters
New York University, US
- Alastair F. Donaldson
Imperial College London, UK
- Juan Pablo Galeotti
Universität des Saarlandes – Saarbrücken, DE
- Vijay Ganesh
University of Waterloo, CA
- Indradeep Ghosh
Fujitsu Labs of America Inc. – Sunnyvale, US
- Arnaud Gotlieb
Simula Research Laboratory – Lysaker, NO
- Istvan Haller
Free Univ. of Amsterdam, NL
- Wei Le
Iowa State Univ. – Ames, US
- Paul Marinescu
Imperial College London, UK
- Benjamin Mehne
University of California – Berkeley, US
- Martin Ochoa
TU München – DE
- Albert Oliveras
Polytechnic University of Catalonia – Barcelona, SP
- Hristina Palikareva
Imperial College London, UK
- Ruzica Piskac
Yale University – New Haven, US
- John Regehr
University of Utah – Salt Lake City, US
- Abhik Roychoudhury
National University of Singapore – SG
- Neha Rungta
NASA – Moffett Field, US
- Raimondas Sasnauskas
University of Utah – Salt Lake City, US
- Koushik Sen
University of California – Berkeley, US
- Oscar Soria Dustmann
RWTH Aachen, DE
- Nikolai Tillmann
Microsoft Corporation – Redmond, US
- Willem Visser
Stellenbosch University – Matieland, ZA
- Klaus Wehrle
RWTH Aachen, DE
- Nicky Williams
CEA LIST – Paris, FR
- Christoph M. Wintersteiger
Microsoft Res. – Cambridge, UK
- Lingming Zhang
University of Texas at Dallas, US