DAGSTUHL
REPORTS

Volume 2, Issue 11, November 2012

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/dagrep

Publication date
March, 2013

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`.

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

Digital Object Identifier: 10.4230/DagRep.2.11.i

Report from Dagstuhl Seminar 12451

The Constraint Satisfaction Problem: Complexity and Approximability

Edited by

Johan Håstad[1], Andrei Krokhin[2], and Dániel Marx[3]

1 **KTH Stockholm, SE,** johanh@kth.se
2 **Durham University, GB,** andrei.krokhin@durham.ac.uk
3 **MTA Budapest, HU,** dmarx@cs.bme.hu

Abstract

During the past two decades, an impressive array of diverse methods from several different mathematical fields, including algebra, logic, analysis, probability theory, graph theory, and combinatorics, have been used to analyze both the computational complexity and approximabilty of algorithmic tasks related to the constraint satisfaction problem (CSP), as well as the applicability/limitations of algorithmic techniques. The Dagstuhl Seminar 12451 "The Constraint Satisfaction Problem: Complexity and Approximability" was aimed at bringing together researchers using all the different techniques in the study of the CSP, so that they can share their insights. This report documents the material presented during the course of the seminar.

Seminar 4.–9. November, 2012 – www.dagstuhl.de/12451
1998 ACM Subject Classification F.2.0 Analysis of Algorithms and Problem Complexity, General
Keywords and phrases Constraint satisfaction problem (CSP), Computational complexity, CSP dichotomy conjecture, Hardness of approximation, Unique games conjecture, Fixed-parameter tractability, Descriptive complexity, Universal algebra, Logic, Decomposition methods
Digital Object Identifier 10.4230/DagRep.2.11.1
Edited in cooperation with Anna Huber

1 Executive Summary

Johan Håstad
Andrei Krokhin
Dániel Marx

The *constraint satisfaction problem*, or CSP in short, provides a unifying framework in which it is possible to express, in a natural way, a wide variety of computational problems dealing with mappings and assignments, including satisfiability, graph colorability, and systems of equations. The CSP framework originated 25-30 years ago independently in artificial intelligence, database theory, and graph theory, under three different guises, and it was realised only in the late 1990s that these are in fact different faces of the same fundamental problem. Nowadays, the CSP is extensively used in theoretical computer science, being a mathematical object with very rich structure that provides an excellent laboratory both for classification methods and for algorithmic techniques, while in AI and more applied areas of computer science this framework is widely regarded as a versatile and efficient way of

The CSP: Complexity and Approximability, *Dagstuhl Reports*, Vol. 2, Issue 11, pp. 1–19
Editors: Johan Håstad, Andrei Krokhin, and Dániel Marx
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

modelling and solving a variety of real-world problems, such as planning and scheduling, software verification and natural language comprehension, to name just a few. An instance of CSP consists of a set of variables, a set of values for the variables, and a set of constraints that restrict the combinations of values that certain subsets of variables may take. Given such an instance, the possible questions include (a) deciding whether there is an assignment of values to the variables so that every constraint is satisfied, or optimising such assignments in various ways, or (b) finding an assignment satisfying as many constraints as possible. There are many important modifications and extensions of this basic framework, e.g. those that deal with soft or global constraints.

Constraint satisfaction has always played a central role in computational complexity theory; appropriate versions of CSPs are classical complete problems for most standard complexity classes. CSPs constitute a very rich and yet sufficiently manageable class of problems to give a good perspective on general computational phenomena. For instance, they help to understand which mathematical properties make a computational problem tractable (in a wide sense, e.g. polynomial-time solvable or non-trivially approximable, fixed-parameter tractable or definable in a weak logic). It is only natural that CSPs play a role in many high-profile conjectures in complexity theory, exemplified by the Dichotomy Conjecture of Feder and Vardi and the Unique Games Conjecture of Khot.

The recent flurry of activity on the topic of the seminar is witnessed by two previous Dagstuhl seminars, titled "Complexity of constraints" (06401) and "The CSP: complexity and approximability" (09441), that were held in 2006 and 2009, respectively. This seminar was a follow-up to the 2009 seminar. Indeed, the exchange of ideas at the 2009 seminar has led to new ambitious research projects and to establishing regular communications channels, and there is a clear potential of a further systematic interaction that will keep on cross-fertilizing the areas and opening new research directions. The 2012 seminar brought together forty four researchers from different highly advanced areas of constraint satisfaction and involved many specialists who use universal-algebraic, combinatorial, geometric and probabilistic techniques to study CSP-related algorithmic problems.

The seminar included two substantial tutorials: one on the classification of the complexity of constraint languages via methods of logic and universal algebra (given by A. Krokhin from Durham U, UK and R. Willard from Waterloo U, CA), and the other on the approximability of CSP (given by P. Austrin from KTH Stockholm, SE). Other participants presented, in 28 further talks, their recent results on a number of important questions concerning the topic of the seminar.

Concluding Remarks and future plans. The seminar was well received as witnessed by the high rate of accepted invitations and the great degree of involvement by the participants. Because of the multitude of impressive results reported during the seminar and the active discussions between researchers with different expertise areas, the organisers regard this seminar as a great success. With steadily increasing interactions between such researchers, we foresee a new seminar focussing on the interplay between different approaches to studying the complexity and approximability of the CSP. Finally, the organisers wish to express their gratitude to the Scientific Directors of the Dagstuhl Centre for their support of the seminar.

Description of the Topics of the Seminar

Classical computational complexity of CSPs. Despite the provable existence of intermediate (say, between P and NP-complete, assuming P $\neq$ NP) problems, research in computational complexity has produced a widely known informal thesis that "natural problems are almost always complete for standard complexity classes”. CSPs have been actively used to support and refine this thesis. More precisely, several restricted forms of CSP have been investigated in depth. One of the main types of restrictions is the *constraint language* restriction, i.e., a restriction on the available types of constraints. By choosing an appropriate constraint language, one can obtain many well-known computational problems from graph theory, logic, and algebra. The study of the constraint language restriction is driven by the CSP *Dichotomy Conjecture* of Feder and Vardi which states that, for each fixed constraint language, the corresponding CSP is either in P or NP-complete. There are similar dichotomy conjectures concerning other complexity classes (e.g. L and NL). Recent breakthroughs in the complexity of CSP have been made possible by the introduction of the universal-algebraic approach, which extracts algebraic structure from the constraint language and uses it to analyse problem instances. McKenzie’s talk surveyed classes of algebras that arise in this context and Pinsker related this approach with infinite-valued CSPs. The algebraic approach has been applied to prove the Dichotomy Conjecture in many important special cases (e.g. Bulatov’s dichotomy theorems for 3-valued and conservative CSPs), but the general problem remains open. A powerful universal-algebraic theory of absorption has been developed in the last couple of years by Barto and Kozik, specifically motivated by CSP classification questions. This theory has already produced several spectacular classification results resolving long-standing problems (including a characterization of CSPs of bounded width, i.e. solvable by local propagation algorithms), and there is a clear sense that there is much more to come from it. Kozik presented new results on CSPs in NL that are based on the absorption theory.

Algebraic approaches to studying exact exponential and sublinear algorithms for CSPs were presented by Jonsson and Yoshida, respectively.

The complexity of Valued CSPs, which are a significant generalisation of Max CSP, was considered in the talks by Huber, Kolmogorov, Thapper, and Živný. Very strong result were reported, especially the full description of tractable cases by Thapper and Živný. Raghavendra presented results that might lead to closer interchange of ideas between algebraic and probabilistic approaches to CSPs.

The algebraic approach to the complexity of counting solutions for CSPs, with many results, was presented by Bulatov, Dyer, Goldberg, and Jerrum, while Lu reported recent progress on classifying the complexity of Holant problems.

Approximability of CSPs. The use of approximation algorithms is one of the most fruitful approaches to coping with NP-hardness. Hard optimization problems, however, exhibit different behavior with respect to approximability, making it an exciting, and by now, well-developed but far from fully understood, research area. The CSP has always played an important role in the study of approximability. For example, it is well known that the famous PCP theorem has an equivalent reformulation in terms of inapproximability of a certain CSP; moreover, the recent combinatorial proof of this theorem by Dinur in 2006 deals entirely with CSPs. The first optimal inapproximability results by Håstad in 2001 were about certain CSPs, and they led to the study of a new hardness notion called *approximation resistance* (which, intuitively, means that a problem cannot be approximated beyond the approximation ratio given by picking an assignment uniformly at random, even on almost satisfiable instances). Many CSPs have been classified as to whether they are approximation resistant but there is not even a reasonable conjecture for a full classification. Håstad, Huang, and K. Makarychev presented new results on approximation resistance.

In a related development, Guruswami and Zhou have discussed, in 2010, a “hybrid” form of tractability for CSPs, where classical tractability is combined with good approximability on almost satisfiable instances, and they conjecture that CSPs of bounded width have this desirable property. This conjecture was proved by Barto and Kozik in 2012 (and presented by Barto at the seminar), with further results in this direction presented by Dalmau.

Arguably, the most exciting development in approximability in the past five to six years is the work around the *unique games conjecture* (UGC), which was introduced by Khot in 2002. It states that, for CSPs with a certain constraint language over a large enough domain, it is NP-hard to tell almost satisfiable instances from those where only a small fraction of constraints can be satisfied. This conjecture (if true) is known to imply optimal inapproximability results for many classical optimization problems. Moreover, if the UGC is true then, as shown by Raghavendra in 2008, a simple algorithm based on semidefinite programming provides the best possible approximation for all CSPs (though the exact quality of this approximation is unknown). In 2010, Arora *et al.* gave a sub-exponential time algorithm for unique games CSPs, which is based on a new graph decomposition method. This does not give strong evidence in favor or against the conjecture, but it shows that there are important new algorithmic ideas to be discovered. Y. Makarychev presented an asymptotically optimal (modulo UGC) approximation algorithm for the general Max CSP.

Parameterized complexity of CSPs. A different way to cope with NP-hardness is provided by parameterized complexity, which relaxes the notion of tractability as polynomial-time solvability to allow non-polynomial dependence on certain problem-specific parameters. A whole new set of interesting questions arises if we look at CSPs from this point of view. Most CSP dichotomy questions can be revisited by defining a parameterized version; so far, very little work was done in this direction compared to investigations in classical complexity. Interestingly, some of the most tantalizing open problems in parameterized algorithmics (e.g. the fixed-parameter tractability of the BICLIQUE problem) are directly related to complexity of CSPs, and Marx's talk contained an overview of such problems. A new research direction (often called "parameterizing above the guaranteed tight bound") led to unexpected positive results for Max r-SAT by Alon *et al.* in 2010. In this direction, the basic question is to decide the fixed-parameter tractability of the following type of problems: if we know that a random assignment satisfies at least E clauses/constraints in expectation (and hence such an assignment is easy to find), find an assignment that satisfies at least $E+k$ clauses/constraints. Gutin presented recent results in this direction.

Along with the constraint language restriction, another important restriction of CSPs that has been thoroughly investigated is the *structural* restriction, where the way in which the immediate interaction between variables in instances is restricted. In this direction, the notions of (hyper)graph decompositions and treewidth turned out to be particularly important. These notions are core concepts of parameterized algorithmics, and so, it is not surprising that parameterized complexity is an important tool in characterizing structural restrictions that lead to tractable CSPs. In particular, many known classification results with respect to classical complexity in this direction (e.g. Grohe, 2007) use tools from parameterized complexity. Scarcello and Szeider described their new results in this direction.

Logic and the complexity of CSP. Starting from earlier work by Kolaitis ad Vardi, concepts and techniques from logic have provided unifying explanations for many tractable CSPs. This has led to the pursuit of classifications of CSP with respect to *descriptive complexity*, i.e. definability in a given logic. Logics considered in this context include first order logic and its extensions, finite-variable logics, the logic programming language Datalog and its fragments. Kozik's talk described a contribution in this direction.

The CSP can be recast as the problem of deciding satisfiability of existential conjunctive formulas. Natural extensions of this framework that allow counting or universal quantifiers were considered in the talks by Martin and Chen, respectively. Atserias' talk related proof complexity, CSPs, and semidefinite programming.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 Semi-Algebraic Proofs, Gaussian Elimination, and CSPs with Short Proofs of Unsatisfiability

Albert Atserias (UPC – Barcelona, ES)

Despite impressive recent progress in obtaining conditional results, one of the big remaining mysteries is why semi-definite programming appears to be the optimal polynomial-time algorithm for approximating constraint satisfaction problems. The lack of a complete understanding is illustrated by the fact that a small generalization of semi-definite programming, the low-degree sum-of-squares method, remains still a candidate algorithm that could beat the UG-optimal Goemans-Williamson bound for max-cut. This raises the obvious question: how powerful low-degree sum-of-squares methods, or more generally low-degree semi-algebraic proofs, really are? A first observation we offer is that low-degree semi-algebraic dag-like proofs, unlike their tree-like versions, are able to simulate both Gaussian elimination over prime fields and bounded-width constraint propagation. Time permitting, we put the question in the more general context of characterizing which CSPs have polynomial-size proofs of unsatisfiability in a given proof system, and offer a general decidable criterion that, unfortunately, so far applies only to width-1 or linear programming.

3.2 Approximability of Constraint Satisfaction Problems: A Tutorial

Per Austrin (KTH Stockholm, SE)

This talk is intended to be a tutorial on the approximability of constraint satisfaction problems, where the goal is to find an assignment satisfying as many constraints as possible.

The tutorial will cover:

1. use of linear and semidefinite programming to obtain non-trivial approximation guarantees
2. the PCP theorem and the Unique Games Conjecture, and the ideas that allow us to derive optimal hardness of approximation results from them
3. qualitative notions of approximability such as approximation resistance and robust approximation

3.3 Robust Satisfiability of Constraint Satisfaction Problems

Libor Barto (Charles University – Prague, CZ)

Joint work of Barto, Libor; Kozik, Marcin
Main reference L. Barto, M. Kozik, "Robust satisfiability of constraint satisfaction problems," in Proc. of the 44th Symp. on Theory of Computing, STOC'12, pp. 931–940, ACM, 2012.
URL http://dx.doi.org/10.1145/2213977.2214061

An algorithm for a constraint satisfaction problem is called robust if it outputs an assignment satisfying almost all constraints given an almost satisfiable instance. Guruswami and Zhou conjectured that CSP over a fixed constraint language admits and efficient robust algorithm if and only if the CSP has bounded width. We confirm their conjecture. The proof is based on an interesting connection between semidefinite programming relaxations and Prague strategies.

3.4 Counting CSPs and Datalog Fixed Points

Andrei A. Bulatov (Simon Fraser University – Burnaby, CA)

Joint work of Bulatov, Andrei A.; Dalmau, Victor; Thurley, Marc

The problem of counting independent sets in bipartite graphs (#BIS) is important in the study of the approximation complexity of counting CSPs. In their 2003 paper Dyer et al. proved that one of the problems interreducible (with respect to approximation preserving reductions) with #BIS is the problem of counting fixed points of a linear Datalog program. We try to determine for which relational structures B the problem #CSP(B) can be expressed as the problem of finding the number of fixed points.

3.5 Meditations on Quantified Constraint Satisfaction

Hubie Chen (Universidad del País Vasco – Donostia, ES)

Main reference H. Chen, "Meditations on Quantified Constraint Satisfaction," in R.L. Constable, A. Silva, (eds.), "Logic and Program Semantics", LNCS 7230, pp. 35–49, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-29485-3_4
URL http://arxiv.org/abs/1201.6306

The quantified constraint satisfaction problem (QCSP) is the generalization of the CSP where universal quantification is permitted in addition to existential quantification: one is given a structure and a sentence built from atoms, conjunction, and the two quantifiers, and the problem is to decide if the sentence holds on the structure. As with the CSP, one obtains a family of problems by defining, for each structure B, the problem QCSP(B) to be the QCSP where the structure is fixed to be B.

We discuss the research program of trying to classify the complexity of QCSP(B) for each finite structure B. We overview the use of universal-algebraic notions and techniques in this research program, and present conjectures concerning when various complexity behaviors occur. We attempt to emphasize open issues and potential research directions, and promise to present concrete open questions.

A protagonist of the talk is the growth rate of the number of elements needed to generate the powers $A^1, A^2, \ldots$ of an algebra A: showing an at-most polynomial growth rate can (essentially) be translated to a complexity upper bound for a structure with algebra A.

3.6 Robust Approximability with Polynomial Loss.

Victor Dalmau (Univ. Pompeu Fabra – Barcelona, ES)

Joint work of Dalmau, Victor; Krokhin, Andrei
Main reference V. Dalmau, A. Krokhin, "Robust Satisfiability for CSPs: Hardness and Algorithmic Results," submitted for journal publication.
URL http://www.dur.ac.uk/andrei.krokhin/papers/robust1.pdf

An algorithm for a constraint satisfaction problem, CSP(H), is called robust if it outputs an assignment satisfying at least a $(1 - f(\epsilon))$-fraction of constraints for each $(1 - \epsilon)$-satisfiable instance (i.e, such that at most a ϵ-fraction of constraints needs to be removed to make the instance satisfiable), where $f(\epsilon) \to 0$ as $\epsilon \to 0$. Barto and Kozik have shown that CSP(H) admits a robust polynomial algorithm if and only if H has bounded width, confirming a conjecture of Guruswami and Zhou. In the present talk we shall describe some additional requirements that guarantee that CSP(H) has a robust algorithm with polynomial loss, namely, such that $f(\epsilon) = O(\epsilon^{1/k})$ for some k.

3.7 The Complexity of Approximating Conservative Counting CSPs

Martin Dyer (University of Leeds, GB)

Joint work of Dyer, Martin; Chen, Xi; Goldberg, Leslie Ann; Jerrum, Mark; Lu, Pinyan; McQuillan, Colin; Richerby, David

We consider the complexity of approximation for the weighted counting constraint satisfaction problem #CSP(F). We study #CSP(F) in the conservative case, where F contains all unary functions. A classification was known for the Boolean domain, which we extend to problems with general finite domain. If F has a property called weak log-modularity, we show that #CSP(F) is in FP. Otherwise, #CSP(F) is as hard to approximate as #BIS. This is the problem of counting independent sets in a bipartite graph, and is believed to be intractable. We further classify the #BIS-hard problems. If F has a property called weak log-supermodularity, we show that #CSP(F) is as easy as Boolean log-supermodular weighted #CSP. Otherwise, #CSP(F) is NP-hard to approximate. Finally, we show that there is a trichotomy for the binary case. Then #CSP(F) is either in FP, or is equivalent to #BIS, or is NP-hard to approximate.

3.8 Approximate Weighted Boolean #CSPs

Leslie Ann Goldberg (University of Liverpool, GB)

Joint work of Bulatov, Andrei; Dyer, Martin; Goldberg, Leslie Ann; Jerrum, Mark; McQuillan, Colin
Main reference A.A. Bulatov, M. Dyer, L.A. Goldberg, M. Jerrum, C. McQuillan, "The expressibility of functions on the Boolean domain, with applications to Counting CSPs," arXiv:1108.5288v4 [cs.CC].
URL http://arxiv.org/abs/1108.5288

Motivated by a desire to understand the computational complexity of (weighted) counting CSPs, we have developed a notion of functional clones (analogous to relational clones) and have studied the landscape of these clones. This was described in Mark Jerrum's talk. In this talk, we give the applications to weighted counting CSPs. We give a complexity classification for the case in which constraints are functions from Boolean tuples to efficiently-computable non-negative reals. For every finite set F of constraint functions, we show that either (1) Approximate counting CSPs are tractable when constraints are taken from F and from any finite set of unary constraint functions, or (2) There is a finite set of unary constraint functions for which this approximation is difficult (subject to complexity-theoretic assumptions which will be described). If there is a function in F which is not log-supermodular, then the approximation problem is NP-hard (without any additional assumptions).

3.9 CSPs Parameterized Above Tight Lower Bounds

Gregory Z. Gutin (RHUL – London, GB)

Results on Max CSPs such as MaxSat and MaxLin2 and Max Permutation CSPs parameterized above tight lower bounds, were overviewed. A proof that Max-r-Sat parameterized above average is fixed-parameter tractable was presented.

For more information, see, e.g., [1, 2, 3].

References

1 N. Alon, G. Gutin, E.J. Kim, S. Szeider, and A. Yeo, Solving MAX-k-SAT Above a Tight Lower Bound. *Algorithmica*, 61 (2011), 638–655. Preliminary version in *ACM-SIAM Symposium on Discrete Algorithms (SODA) 2010*, pp. 511–517.

2 R. Crowston, M. Fellows, G. Gutin, M. Jones, F. Rosamond, S. Thomassé and A. Yeo, Simultaneously Satisfying Linear Equations Over $\mathbb{F}_2$: MaxLin2 and Max-r-Lin2 Parameterized Above Average. In *FSTTCS 2011*, LIPICS Vol. 13, 229–240.

3 R. Crowston, G. Gutin, M. Jones, V. Raman and S. Saurabh, Parameterized Complexity of MaxSat Above Average. *Theor. Comput. Sci.*, to appear. Pleliminary version in *LATIN 2012*, Lect. Notes Comput. Sci. 7256 (2012), 184–194.

3.10 On the NP-hardness of Max-Not-2

Johan Håstad (KTH Stockholm, SE)

We prove that, for any $\epsilon > 0$, it is NP-hard to, given a satisfiable instance of Max-NTW (Not-2), find an assignment that satisfies a fraction $\frac{5}{8} + \epsilon$ of the constraints. This, up to the existence of ϵ, matches the approximation ratio obtained by the trivial algorithm that just pick an assignment at random and thus the result proves that Max-NTW is approximation resistant on satisfiable instances. This result makes our understanding of arity three Max-CSPs with regards to approximation resistance complete.

3.11 Approximation Resistance on Satisfiable Instances for Predicates with Few Accepting Assignments

Sangxia Huang (KTH – Stockholm, SE)

In this talk we consider approximability of Max-CSP(P), where P is a Boolean predicate of arity k, and the goal is to find an assignment satisfying as many clauses as possible. A predicate is said to be approximation resistant if it is hard to achieve better approximation ratio than random assignment. A related problem is approximation resistance on satisfiable instances, where we are given satisfiable instances. The situation could be quite different in these cases, for instance PARITY (LinearEquation on k variables) is approximation resistant but on satisfiable instances we can solve optimally using Gaussian Elimination.

While there has been lots of progress in understanding approximation resistance of predicates and many tight upper- and lower-bounds are known, approximation resistance on satisfiable instances is still largely a mystery. In this talk, we will survey some known results on this problem, the role of PARITY, and present a construction which proves that some predicate on k variables with $2^{\tilde{O}(k^{1/3})}$ accepting assignments is approximation resistant on satisfiable instances, improving the previous bound of $2^{O(k^{1/2})}$ by (Håstad and Khot, 2001).

3.12 VCSPs on Three Elements

Anna Huber (University of Durham, GB)

Joint work of Huber, Anna; Krokhin, Andrei; Powell, Robert
Main reference A. Huber, A. Krokhin, R. Powell, "Skew Bisubmodularity and Valued CSPs," in Proc. of the 24th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA), pp. 1296–1305, 2013.
URL http://knowledgecenter.siam.org/0236-000082

An instance of the Finite-Valued Constraint Satisfaction Problem (VCSP) is given by a finite set of variables, a finite domain of values, and a sum of rational-valued functions, each function depending on a subset of the variables. The goal is to find an assignment of values to the variables that minimises the sum.

This talk investigates VCSPs in the case when the variables can take three values and provides a tight description of the tractable cases.

3.13 Functional Clones

Mark Jerrum (Queen Mary University of London, GB)

Joint work of Bulatov, Andrei A.; Dyer, Martin; Goldberg, Leslie Ann; Jerrum, Mark; McQuillan, Colin
Main reference A.A. Bulatov, M. Dyer, L.A. Goldberg, M. Jerrum, C. McQuillan, "The expressibility of functions on the Boolean domain, with applications to Counting CSPs," arXiv:1108.5288v4 [cs.CC].
URL http://arxiv.org/abs/1108.5288

In the classical setting, where CSPs model decision problems, a certain notion of expressibility, namely pp-definablity, plays an important role, as do relational clones, i.e., sets of relations closed under pp-definability. The corresponding concepts for valued CSPs (VCSPs), which model optimisation problems, have recently received a good deal of attention, and have yielded significant insight. What is the correct notion of expressibility, i.e., analogue of pp-definability, for (weighted) counting CSPs (#CSPs)? Equivalently, how should we define "functional clone"? The answer is not completely straightforward, as some delicate choices have to be made.

After proposing an answer to the above question, I will go on to investigate the structure of the lattice of functional clones in the Boolean, conservative case. In a second talk, Leslie Goldberg will continue this theme by investigating the consequence of these results for the complexity of #CSPs in the Boolean, conservative case.

3.14 Complexity of SAT Problems, Clone Theory and the Exponential Time Hypothesis

Peter Jonsson (Linköping University, SE)

Joint work of Jonsson, Peter; Lagerkvist, Victor; Nordh, Gustav; Zanuttini, Bruno
Main reference P. Jonsson, V. Lagerkvist, G. Nordh, B. Zanuttini, "Complexity of SAT Problems, Clone Theory and the Exponential Time Hypothesis," in Proc. of the 24th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA), pp. 1264–1277, 2013.
URL http://knowledgecenter.siam.org/0236-000094

The construction of exact exponential-time algorithms for NP-complete problems has for some time been a very active research area. Unfortunately, there is a lack of general methods for studying and comparing the time complexity of algorithms for such problems. We propose a method based on the lattice of partial clones and demonstrate it on the SAT problem. By using this method, we identify a relation R_e such that SAT(R_e) is, in a certain sense, the computationally easiest NP-complete SAT problem. We additionally demonstrate that SAT(R_e)-2 (i.e. SAT(R) restricted to instances where no variable appears in more than two clauses) is NP-complete, too. We then relate SAT(R_e)-2 to the exponential-time hypothesis (ETH) and show that ETH holds if and only if SAT(R)-2 is not sub-exponential. This constitutes a strong connection between ETH and the SAT problem under both severe relational and severe structural restrictions. In the process, we also prove a stronger version of Impagliazzo et. al's sparsification lemma for k-SAT; namely that all finite, NP-complete

Boolean languages can be sparsified into each other. This should be compared with Santhanam and Srinivasan's recent negative result which states that the same does not hold for all infinite Boolean languages.

3.15 Linear Programming and VCSPs Revisited

Vladimir Kolmogorov (IST Austria – Klosterneuburg, AT)

Main reference V. Kolmogorov, "The power of linear programming for valued CSPs: a constructive characterization," arXiv:1207.7213v4 [cs.CC]
URL http://arxiv.org/abs/1207.7213

I study which classes of finite-valued VCSPs can be solved exactly by the Basic Linear Programming relaxation (BLP). Thapper and Zivny proved that BLP solves a language iff it admits a symmetric fractional polymorphism of every arity. I show that it's sufficient to have a *binary* symmetric fractional polymorphism.

Combined with the recent dichotomy result of Thapper and Zivny, this implies that a finite-valued language can either be solved by BLP or it is NP-hard.

Link to the paper: http://pub.ist.ac.at/~vnk/papers/BLP.html

3.16 Some of the CSP's Solvable in Linear Datalog

Marcin Kozik (Jagiellonian University – Kraków, PL)

Joint work of Kozik, Marcin; Barto, Libor; Willard, Ross

Larose and Tesson conjectured that a complement of a CSP with the template in a variety omitting types 1,2 and 5 is definable in linear datalog. The result has been proved for templates with majority polymorphism (Dalmau, Krokhin) and near-unanimity polymorphism (Barto, Kozik, Willard). I will discuss the last proof and some further developments.

3.17 A Tutorial on Algebra and CSP, Part 1

Andrei Krokhin (University of Durham, GB)

I will explain why algebra works in classifying the complexity of CSPs and how algebraic classification results underpin complexity classification results. The second part of tutorial, given by Ross Willard, will be devoted to showing how algebra is used to design and analyse algorithms for CSPs.

3.18 Holant Problems: CSPs Where Each Variable Appears Exactly Twice

Pin-Yan Lu (Microsoft Research Asia, CN)

Holant Problem is a general framework to capture local constraints. CSP can be viewed as a special family of Holant problems where equality constraints of all arities are assumed to be available. This framework allows for the expression of (perfect) matching problems, a class of substantive combinatorial problems that have proved pivotal in complexity theory. Dichotomy is still open in this framework even for the Boolean domain case both for decision version and counting version. In this talk, I will review some of the results in this framework with an emphasis on some new phenomena comparing to the CSP framework. I will talk about decision version (NP), counting version (#P) and parity version ($\oplus$ P) with a focus on the counting version.

3.19 Local Search is Better than Random Assignment for Bounded Occurrence Ordering k-CSPs

Konstantin Makarychev (Microsoft Research – Redmond, US)

We prove that the Bounded Occurrence Ordering k-CSP Problem is not approximation resistant. We give a very simple local search algorithm that always performs better than the random assignment algorithm. Specifically, the expected value of the solution returned by the algorithm is at least Alg > Avg + $a(B, k)$ (Opt - Avg), where "Opt" is the value of the optimal solution; "Avg" is the expected value of the random solution; and $a(B, k) = \Omega_k(B^{-(k+O(1))}$ is a parameter depending only on "k" (the arity of the CSP) and "B" (the maximum number of times each variable is used in constraints).

The question whether bounded occurrence ordering k-CSPs are approximation resistant was raised by Guruswami and Zhou (APPROX 2012) who recently showed that bounded occurrence 3-CSPs and "monotone" k-CSPs admit a non-trivial approximation.

3.20 Approximation Algorithm for Non-Boolean MAX k-CSP

Yury Makarychev (Toyota Technological Institute – Chicago, US)

Joint work of Makarychev, Konstantin; Makarychev, Yury
Main reference K. Makarychev, Y. Makarychev, "Approximation Algorithm for Non-boolean MAX k-CSP. Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques," in Proc. of 15th Int'l APPROX Workshop and 16th Int'l RANDOM Workshop, LNCS, Vol. 7408, pp. 254–265, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-32512-0_22

We present a randomized polynomial-time approximation algorithm for Max k-CSP$_d$. In Max k-CSP$_d$, we are given a set of predicates of arity k over an alphabet of size d. Our goal is to find an assignment that maximizes the number of satisfied constraints. Our algorithm has approximation factor $\Omega(kd/d^k)$ (when $k \geq \Omega(\log d)$). The best previously known algorithm has approximation factor $\Omega(k \log d/d^k)$.

Our bound is asymptotically optimal. We also give an approximation algorithm for the boolean Max k-CSP$_2$ problem with a slightly improved approximation guarantee.

3.21 Constraint Satisfaction with Counting Quantifiers

Barnaby Martin (Middlesex University, GB)

We consider CSPs and QCSPs augmented with counting quantifiers.

3.22 CSPs and Fixed-Parameter Tractability

Daniel Marx (MTA – Budapest, HU)

We survey fixed-parameter tractability results appearing in the context of constraint satisfaction. The focus of the talks in on explaining the different type of questions that can be asked and the briefly summarizing the known results without going into the technical details.

3.23 The Dichotomy Conjecture: Sketching the Algebraic Landscape Near the Boundary

Ralph McKenzie (Vanderbilt University – Nashville, US)

I shall look at the chief classes of finite algebras that have assumed importance in the algebraic attempts to resolve the dichotomy conjecture, and survey many of the results that have emerged, sketching their significance for constraint satisfaction, and as contributions to pure algebra.

3.24 Topological Birkhoff

Michael Pinsker (University Paris-Diderot, FR)

Joint work of Pinsker, Michael; Bodirsky, Manuel

I will present a new method for hardness proofs of CSPs with (infinite) omega-categorical templates: we prove that if S is an omega-categorical countable structure whose polymorphism clone allows a continuous homomorphism to the trivial clone of projections, then CSP(S) is NP-hard.

The proof of this statement is based on a generalization of the finite version of Birkhoff's HSP theorem to omega-categorical structures, and I will outline the connection with this theorem. Moreover, I will describe the larger program in which we wish to reduce CSPs of infinite templates to CSPs of finite templates, and of which the above-mentioned result is an important step.

3.25 Efficient Algorithms via Polymorphisms in the Value Oracle Model

Prasad Raghavendra (University of California – Berkeley, US)

In this work, we design efficient algorithms via fractional polymorphisms for problems that are not specified as constraint satisfaction problems. Specifically, we show the following result:

1) Suppose we are given access to a function F in the value oracle model, along with an operation/polymorphism on the domain that never increases the value of F. Submodular minimization is a well-known example of this nature, with the operation being 2-bit AND and 2-bit OR.

We show that under restrictions on the operation/polymorphism, the function F can be minimized in pseudopolynomial time. This shows that the tractability of submodular minimization in the value-oracle model is a special case of a general phenomenon.

3.26 Tree Projections and Structural Decomposition Methods: The Power of Local Consistency and Larger Islands of Tractability

Francesco Scarcello (University of Calabria, IT)

Joint work of Greco, Gianluigi; Scarcello, Francesco
Main reference G. Greco, F. Scarcello, "Tree Projections and Structural Decomposition Methods: The Power of Local Consistency and Larger Islands of Tractability," arXiv:1205.3321v2 [cs.DB], 2012. (Preliminary version in the Proceedings of PODS'10.)
URL http://arxiv.org/abs/1205.3321

Evaluating conjunctive queries and solving constraint satisfaction problems are fundamental problems in database theory and artificial intelligence, respectively. These problems are NP-hard, so that several research efforts have been made in the literature for identifying tractable classes, known as islands of tractability, as well as for devising clever heuristics for solving efficiently real-world instances. Many heuristic approaches are based on enforcing on the given instance a property called local consistency, where (in database terms) each tuple in every query atom matches at least one tuple in every other query atom. Interestingly, it turns out that, for many well-known classes of instances, such as the acyclic ones, enforcing local consistency is even sufficient to solve the given instance correctly. However, the precise power of such a procedure was unclear, but for some very restricted cases. We provide the answers to the long-standing questions about the precise power of algorithms based on enforcing local consistency. The classes of instances where enforcing local consistency turns out to be a correct CSP-solving procedure are however not efficiently recognizable. In fact, the paper finally focuses on certain subclasses defined in terms of the novel notion of greedy tree projections. These latter classes are shown to be efficiently recognizable and strictly larger than most islands of tractability known so far, both in the general case of tree projections and for specific structural decomposition methods.

3.27 Structural Parameterizations of Language Restricted Constraint Satisfaction Problems

Stefan Szeider (Vienna University of Technology, AT)

Joint work of Szeider, Stefan; Bova, Simone

We study the fixed-parameter tractability of the constraint satisfaction problem. We restrict constraint relations to languages in a family of NP-hard languages, classified by a purely combinatorial criterion that generalizes Boolean matrices with fixed row and column sum. For various natural and established structural parameterizations of the instances, we characterize the fixed-parameter tractable constraint languages in the family.

3.28 The Complexity of Finite-Valued CSPs

Johan Thapper (Ecole Polytechnique – Palaiseau, FR)

Joint work of Thapper, Johan; Zivny, Stanislav
Main reference J. Thapper, S. Zivny, "The complexity of finite-valued CSPs," arXiv:1210.2987v2 [cs.CC].
URL http://arxiv.org/abs/1210.2987

I will give a complete complexity classification for finite-valued CSPs. The final result states that every core language Γ either admits a binary idempotent and symmetric fractional polymorphism, in which case the basic linear programming relaxation solves VCSP(Γ) exactly, or Γ satisfies a simple hardness condition that allows for a polynomial-time reduction from Max-Cut.

3.29 A Tutorial on Algebra and CSP, Part 2

Ross Willard (University of Waterloo, CA)

In this talk I explained (roughly) the two currently most-general polynomial-time CSP algorithms ("local consistency" and "few subpowers"), indicating in particular the role of algebra and polymorphisms.

3.30 Algebraic Characterizations of Testable CSPs

Yuichi Yoshida (NII – Tokyo, JP)

Joint work of Yoshida, Yuichi; Arnab Bhattacharyya
Main reference A. Bhattacharyya, Y. Yoshida, "Testing Assignments of Boolean CSPs," ECCC TR12–103, 2012.
URL http://eccc.hpi-web.de/report/2012/103/

Given an instance I of a CSP, a tester for I distinguishes assignments satisfying I from those which are far from any assignment satisfying I. The efficiency of a tester is measured by its query complexity, the number of variable assignments queried by the algorithm. In this talk, we show a characterization of the hardness of testing Boolean CSPs in terms of the associated algebra. In terms of computational complexity, we show that if a non-trivial Boolean CSP is sublinear- query testable (resp., not sublinear-query testable), then the CSP is in NL (resp., P-complete, parityL-complete or NP-complete) and that if a sublinear-query testable Boolean CSP is constant-query testable (resp., not constant-query testable), then counting the number of solutions of the CSP is in P (resp., #P- complete). Additionally, we conjecture that a CSP instance is testable in sublinear time if its Gaifman graph has bounded treewidth. We confirm the conjecture when a near- unanimity operation is a polymorphism of the CSP. We also mention a similar characterization for list H-homomorphism problem.

3.31 Linear Programming and VCSPs

Stanislav Zivny (University of Warwick, GB)

Joint work of Thapper, Johan; Zivny, Stanislav
Main reference J. Thapper, S. Zivny, "The Power of Linear Programming for Valued CSPs," in Proc. of the 53rd Annual IEEE Symp. on Foundations of Computer Science (FOCS'12), pp. 669–678, IEEE, 2012.
URL http://dx.doi.org/10.1109/FOCS.2012.25

This talk presents a recent result on the power of LP for valued CSPs: a valued constraint language L is solvable by the basic linear programming relaxation (BLP) iff L admits symmetric fractional polymorphisms of all arities (and 2 more equivalent statements).

Our results establish tractability of several previously widely-open classes of VCSPs including (i) VCSPs that are submodular on *arbitrary* lattices, (ii) VCSPs that are bisubmodular (also known as k-submodular) on *arbitrary* domains, and (iii) VCSPs that are submodular on *arbitrary* trees.

Participants

- Albert Atserias
UPC – Barcelona, ES
- Per Austrin
KTH Stockholm, SE
- Libor Barto
Charles University – Prague, CZ
- Manuel Bodirsky
Ecole Polytechnique – Palaiseau, FR
- Andrei A. Bulatov
Simon Fraser University – Burnaby, CA
- Catarina Carvalho
University of Hertfordshire, GB
- Hubie Chen
Universidad del País Vasco – Donostia, ES
- David Cohen
RHUL – London, GB
- Victor Dalmau
Univ. Pompeu Fabra – Barcelona, ES
- Martin Dyer
University of Leeds, GB
- Leslie Ann Goldberg
University of Liverpool, GB
- Martin Grohe
RWTH Aachen, DE
- Venkatesan Guruswami
Carnegie Mellon University – Pittsburgh, US
- Gregory Z. Gutin
RHUL – London, GB
- Johan Hastad
KTH Stockholm, SE
- Sangxia Huang
KTH – Stockholm, SE
- Anna Huber
University of Durham, GB
- Mark Jerrum
Queen Mary University of London, GB
- Peter Jonsson
Linköping University, SE
- Vladimir Kolmogorov
IST Austria – Klosterneuburg, AT
- Marcin Kozik
Jagiellonian Univ. – Kraków, PL
- Stefan Kratsch
TU Berlin, DE
- Andrei Krokhin
University of Durham, GB
- Benoit Larose
Champlain Regional College – St. Lambert, CA
- Pin-Yan Lu
Microsoft Res. Asia, CN
- Konstantin Makarychev
Microsoft Res. – Redmond, US
- Yury Makarychev
Toyota Technological Institute – Chicago, US
- Petar Markovic
University of Novi Sad, SEU
- Barnaby Martin
Middlesex University, GB
- Dániel Marx
MTA – Budapest, HU
- Ralph McKenzie
Vanderbilt Univ. – Nashville, US
- Michael Pinsker
University Paris-Diderot, FR
- Prasad Raghavendra
University of California – Berkeley, US
- Francesco Scarcello
University of Calabria, IT
- Ola Svensson
EPFL – Lausanne, CH
- Stefan Szeider
Vienna Univ. of Technology, AT
- Suguru Tamaki
Kyoto University, JP
- Johan Thapper
Ecole Polytechnique – Palaiseau, FR
- Matt Valeriote
McMaster Univ. – Hamilton, CA
- Moshe Y. Vardi
Rice University, US
- Magnus Wahlström
MPI für Informatik – Saarbrücken, DE
- Ross Willard
University of Waterloo, CA
- Yuichi Yoshida
NII – Tokyo, JP
- Stanislav Zivny
University of Warwick, GB

Report from Dagstuhl Perspectives Workshop 12452

Publication Culture in Computing Research

Edited by

Kurt Mehlhorn[1], Moshe Y. Vardi[2], and Marc Herbstritt[3]

1 **MPI für Informatik – Saarbrücken, DE,** mehlhorn@mpi-inf.mpg.de
2 **Rice University, US,** vardi@cs.rice.edu
3 **Schloss Dagstuhl – Leibniz-Zentrum für Informatik, DE,** marc.herbstritt@dagstuhl.de

Abstract

The dissemination of research results is an integral part of research and hence a crucial component for any scientific discipline. In the area of computing research, there have been raised concerns recently about its publication culture, most notably by highlighting the high priority of conferences (compared to journals in other disciplines) and – from an economic viewpoint – the costs of preparing and accessing research results.

The Dagstuhl Perspectives Workshop 12452 "Publication Culture in Computing Research" aimed at discussing the main problems with a selected group of researchers and practitioners. The goal was to identify and classify the current problems and to suggest potential remedies. The group of participants was selected in a way such that a wide spectrum of opinions would be presented. This lead to intensive discussions.

The workshop is seen as an important step in the ongoing discussion. As a main result, the main problem roots were identified and potential solutions were discussed. The insights will be part of an upcoming manifesto on Publication Culture in Computing Research.

Seminar 06.–09. November, 2012 – www.dagstuhl.de/12452
1998 ACM Subject Classification K.4 Computer and Society, K.7 The Computing Profession, K.7.2 Organizations, K.7.4 Professional Ethics
Keywords and phrases scholarly publishing, conference, journal, peer review, open archive, open access, indexing, research evaluation
Digital Object Identifier 10.4230/DagRep.2.11.20
Edited in cooperation with Andrew P. Bernat, Jon Crowcroft, Jan van Leeuwen, Bertrand Meyer, Fred B. Schneider, and Douglas B. Terry

1 Executive Summary

Kurt Mehlhorn
Moshe Y. Vardi
Marc Herbstritt

The dissemination of research results is an integral part of research and hence a crucial component for any scientific discipline. While computing research has been phenomenally successful, there is a broad feeling that the publication models are quite often obstacles. Yet there is no agreement on whether the publication models need to be radically changed or fine tuned, and there is no agreement on how such change may occur. Over the past few years, a vigorous

Publication Culture in Computing Research, *Dagstuhl Reports*, Vol. 2, Issue 11, pp. 20–44
Editors: Kurt Mehlhorn, Moshe Y. Vardi, and Marc Herbstritt
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

discussion has been going on through editorials, Viewpoint articles, and blogs of the Communication of the ACM – see Jonathan Grudin's overview available at http://research.microsoft.com/en-us/UM/People/jgrudin/publications/publicationculture/CACMreferences.pdf.

In spite of this ongoing debate, the community seems no closer to an agreement whether a change has to take place and how to effect such a change.

The workshop brought together key players in this debate for an intense three-day discussion and deliberation, with the aim of analyzing the issues and developing guidelines for the way forward. A specific focus of the workshop was to develop consensus around a set of guiding principles. An expected outcome of the workshop is a manifesto to be published afterwards.

Topics

The workshop addressed several topics that were part of the community's collective conversation on publication culture during the last years:

1. The uniqueness of the publication model in computing research:
   - the emphasis on conference publishing and the decline of journal publishing;
   - the large and growing number of specialty conferences and workshops that are really conferences;
   - coping with established publication cultures in the (other) sciences and with the different cultures of different computing sub-communities.
2. Cultural issues:
   - the culture of hypercritical reviewing and the decline of thorough constructive reviewing;
   - tenure and promotion practices that encourage short-term research;
   - the influence of bibliometry on publication behavior and tenure practices and the quality of bibliometry.
3. New publication models:
   - the tension between open access and reader-pays publishing, and the spectrum in between;
   - the role of social media in scholarly publishing;
   - the role of various actors: commercial publishers, scientific societies, academic publishers and archives;
   - the place of self-publishing or publishing in public repositories;
   - the need to develop new rules for data citation, sharing, and archiving.

Organization

The workshop was organized by Moshe Y. Vardi and Kurt Mehlhorn with coordinating support by Marc Herbstritt. Additionally, a program committee (PC) was set up, including Andrew P. Bernat, Jon Crowcroft, Jan van Leeuwen, Bertrand Meyer, Fred B. Schneider, and Douglas B. Terry. The PC helped in seeking suitable contributions and advising the organizers in shaping the program. Each invitee was asked to submit a position statement which was reviewed by the organizers and the PC. The collection of accepted position statements provided a broad and concise overview of the problems in the publication culture of computing research, disclosing a variety of different and competing viewpoints.

On Wednesday Nov. 7, 2012, the workshop started with a session presenting standpoints from scholarly societies and commercial publishers, among them Ronald Boisvert

(NIST/ACM), Dan Wallach (Rice University/USENIX), Maarten Fröhlich (IOS Press), Alfred Hofmann (Springer Science Business+Media/LNCS), Sweitze Roffel (Elsevier), Andrew Bernat (Computing Research Association), and Moshe Y. Vardi (Rice University/ Editor-in-Chief of Comm. of the ACM). The afternoon session focussed on peer review and research dissemination, including the talks from Bertrand Meyer (ETH Zürich/Informatics Europe), Ursula Martin (Queen Mary University London), Lance Fortnow (Georgia Institute of Technology), Doug Terry (Microsoft – Mountain View), Nicolas Holzschuch (INRIA Rhône-Alpes), George Danezis (Microsoft Research – Cambridge), and José Palazzo Moreira de Oliveira (UFRGS).

On Thursday Nov. 8, 2012, the workshop continued with a morning session on "conferences versus journals" as well as on "open access", with talks from Manuel Hermenegildo (IMDEA), Keith Marzullo (NSF), Kurt Mehlhorn (MPII), Jeff Mogul (HP), M. Tamer Özsu (University of Waterloo), and Vladimiro Sassone (University of Southampton). The afternoon session focussed also "conferences versus journals", but also on indexing and general cultural issues; talks were given by Reinhard Wilhelm (Saarland University), Jan van Leeuwen (Utrecht University), Jonathan Grudin (Microsoft Research – Redmond), Andrei Voronkov (Manchester University), Srinivasan Keshav (University of Waterloo), Fred B. Schneider (Cornell University), and Batya Friedman (University of Washington).

Batya Friedman moderated the "Future Workshop", which (1) interactively asked participants after the sessions to contribute brief descriptions of substantial shortcomings in our current publication culture, according to one's own opinion, (2) asked participants to describe an idealized publication culture for computing research, and (3) finally, asked participants to provide brief accounts of potential solutions to the problems raised and ways to reach ideal outcomes.

The results of the "Future Workshop" were discussed on Friday Nov. 9, 2012, and served as basis for working groups. The working groups met in small teams and presented the results from their discussions to the audience. Finally, Moshe Y. Vardi gave a summary on the workshop and talked about future actions.

The organizers and the PC met on Friday afternoon to clarify core issues for the upcoming manifesto.

Outcomes

The main outcomes will be covered in the upcoming manifesto that will be published in the "Dagstuhl Manifestos" series[1]. However, as discussed during the organizers and PC meeting on Friday afternoon, a first sketch of a consensus list with regard to problems and desired solutions is as follows:

- Problems:
  - *Scaling*: The publishing ecosystem in computing research—conference and journals—has not scaled up.
  - *Policy*: We have no universally accepted norms and policies, and no single authority.
  - *Data*: We have many opinions but little data.
  - *Business model*: Huge gap between publishers and authors/readers.

[1] http://drops.dagstuhl.de/dagman

  - *Incentives*: Large number of small papers.
  - *Measurements*: Highly imperfect metrics.
  - *Conferences*: Too many submissions and resubmissions, random-like decisions, too many conferences, too much travel, conferences as “journals that meet in hotels”.
  - *Journals*: Not exciting, hard to find reviewers, poor journal management systems.
  - *Reviewing*: Increasing burden, declining standards, declining competence.
- Wish list:
  - *Defragmented Community*: Learn to operate at scale.
  - *Rational reviewing*: Eliminate treadmill, eliminate hyper-criticality, reintroduce review rounds.
  - *Revitalized journals*: Perhaps through jourference/cournal hybrids.
  - *Reduce paper inflation*: Focus on quality, not quantity.
  - *Appropriate bibliometrics*: Recognize conferences, eliminate self-citation.
  - *Open Access*: Research results should be available to all to read and assess as soon as possible.
  - *Viable associations*: Strong associations that can enable, facilitate, and lead a better publication culture.

Resources

Position statements and slides from the presentations are available at http://www.dagstuhl.de/mat/index.en.phtml?12452.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

Further links, immediate outcome, and follow-up discussions

3 Overview of Talks

3.1 Thoughts on the Publication Culture in Computing Research

Andrew P. Bernat (Computing Research Association, US)

The computing fields have a unique publication culture that relies heavily on conferences as a venue for publishing scholarly work. This situation predates the 1999 CRA Best Practices Memo "Evaluating Computer Scientists and Engineers For Promotion and Tenure" by Patterson, Snyder and Ullman [1], which laid out the reasons for this situation. The intent of that memo was simply to lay out how "impact" should be evaluated in the computing research field, but it has been used ever since as a justification for publishing in conferences and downplaying the importance of publishing in journals, the typical practice in other scientific disciplines.

The response was quick and the discussion has continued to this day, from arguments for a grand-scale rethinking of publication practices to a discussion of the "negative" impacts of our "conferences as journals" rather than "conferences as community".

Adopting a different publication model has not been impact free to the computing research community; among the negative impacts are:

1. the large and growing number of specialty conferences and workshops that are really conferences
2. coping with established publication cultures in the (other) sciences and with the different cultures of different computing sub-communities
3. the culture of hypercritical reviewing and the decline of thorough constructive reviewing
4. tenure and promotion practices that encourage short-term research
5. the influence of bibliometry on publication behavior and tenure practices and the quality of bibliometry
6. by emphasizing conference publication there is an inevitable de-emphasis on the other values inherent in a conference including:
   (a) creating opportunities for cross-fertilization
   (b) creating a community gathering
   (c) providing a venue for high quality dissemination of research results
   (d) providing a high profile public venue for computing research

At the same time, it is clear that the publication model in use today throughout scholarly endeavors is under considerable stress:

1. the tension between open access and reader-pays publishing and the spectrum in between
2. the role of social media in scholarly publishing
3. the role of various actors: commercial publishers, scientific societies, academic publishers and archives
4. the place of self-publishing or publishing in public repositories
5. the need to develop new rules for data citation, sharing, and archiving
6. the impact of new publishing models on the business models of our societies

References

1 D. Patterson, L. Snyder, J. Ullman, Evaluating computer scientists and engineers for promotion and tenure, Best Practice Memo, Computing Research Association, September 1999, `http://cra.org/resources/bp-view/evaluating_computer_scientists_and_engineers_for_promotion_and_tenure/`.

3.2 Fair Access: Society Publishing and the Subscription Model

Ronald F. Boisvert (NIST – Gaithersburg, US)

Joint work of Boisvert, Ronald F.; Davidson, Jack;

We present the perspective of non-profit professional societies as scholarly publishers, which we characterize as "fair access." Fair access is a subscription model based on low cost for access and liberal author rights including self-archiving, with surpluses used to improve the health of the profession. We describe implementation of fair access by the Association for Computing Machinery (ACM), an educational society governed by its members, and how it is continuing to adapt its policies to changes within the computer science community.

3.3 The peer-review process: Insights from the largest computer security conferences

George Danezis (Microsoft Research UK – Cambridge, GB)

URL http://conspicuouschatter.wordpress.com/2011/10/23/acm-ccs-2011-reviewing/

In 2011 and 2012 ACM Computers & Communication Security (CCS) conferences received, and accepted, an unprecedented number of submissions, making them the largest academic computer security conferences to date. In 2011 the surge in submissions was unexpected and the peer review process had to be streamlined to make optimal use of the reviewing resources available. This forced the chairs to have a very close look at the quantitative and qualitative data from the whole process, as well as assess its function in the overall decision making process. Over the two years over 2000 reviews were filed from about 100 PC members and hundreds of external reviewers, making it the largest data set to study the variability of reviews, scores, and their evolution over the course of the review process. In both years program committee members had a large workload, that allows us to study different profiles of reviewers, as well as the variance between individuals. In this talk I will present some tools we developed based on these data-sets, and some conclusions about publishing culture supported by both quantitative and qualitative data.

Related blog post as position statement: http://conspicuouschatter.wordpress.com/2011/10/23/acm-ccs-2011-reviewing/

3.4 Separating the Editorial from the Distribution in CS Publications

Lance Fortnow (Georgia Institute of Technology, US)

Conference and Journal publications in computer science have traditionally played multiple roles: As a method to choose great papers, give comments to the authors and to distribute those publications to the CS community. In many industries, like books, music, movies and newspapers, the Internet has broken this link between content and distribution. While there is no direct mapping from academic publications to those industries, we nevertheless need to rethink the role of publications in this new environment. The academic community, especially in computer science, wants their publications to be accessible to everyone at any time. People should post their research papers on unedited or lightly edited publicly accessible archive sites. In addition there should be a system that allows conferences and journals to choose from these papers, both to help the community find the best papers in the field as well as a method of giving authors a way to claim quality on their papers. This position paper will discuss a few different approaches toward this goal, discussing some of the benefits and challenges of each.

3.5 Intellectual Culture: Perspectives on the Publication Culture in Computing Research

Batya Friedman (University of Washington, US)

In my remarks I will reflect on the current publication and intellectual culture in information and computer science, particularly the ways in which the culture and practices incentivize (though perhaps not intentionally) rapid, small(er), stand-alone, non-interdisciplinary intellectual contributions; and place a downward pressure on young(er) scholars to do more and to achieve more earlier and earlier in their careers. I will discuss the implications of such a culture for developing and valuing mature scholars and deep scholarship. Then I will contrast the current culture with one that prioritizes scholarship that builds substantively on prior work, values longer-term programmatic research with theory construction, supports careful implementation and reporting of method, and appropriately rewards authentic interdisciplinary efforts. As time permits, I will suggest some ideas toward the recovery of mature scholarship. In so doing, I will touch on what making such a cultural transition might entail including ethical issues, and call out potential hazards along such a path.

3.6 Publishing at IOS Press

Maarten Froehlich (IOS Press, NL)

Of the 100 or so journals and approximately 125 books which IOS Press publishes each year, 30 journals and about 75 books fall broadly into the category of Information and Communication Sciences. An increasing number of these are now offered in an Open Access context, both for our journal and book platforms. These open access publications are referred to as the IOS Press Open Library®. Open peer review is another area which is also being explored successfully. We offer services to institutional scientists, organisations and conferences using a range of publishing models, from full e-publishing to e and paper. The latter is still in demand for a variety of purposes, but everything we publish exists in digital form as part of the production workflow. Clients and authors decide on how they want the material to be delivered to them.

3.7 Conferences, Journals, and the Competitive Exclusion Principle

Jonathan Grudin (Microsoft Research – Redmond, US)

Journals were invented to serve purposes not served by meetings, notably to widely disseminate and permanently archive results. Conferences served to build and maintain the community that comprises a field. For centuries, journals and conferences occupied these distinct niches. In computer science today, conference proceedings are widely disseminated and permanently archived. In these and other ways, conferences have invaded the niche occupied by journals. The competitive exclusion principle formulated by evolutionary biologists states that two species cannot occupy the same niche: One will become extinct or be forced to move to a different niche. We have a third option, rarely possible in biology – we can create a single hybrid. I will describe numerous ongoing experiments to create conference-journal hybrids or mergers. I argue that the conference-journal question will likely work itself out, and that a more critical issue is the empty community-building niche vacated by conferences.

3.8 Conferences vs. Journals in CS, what to do? Evolutionary ways forward and the ICLP/TPLP Model

Manuel Hermenegildo (IMDEA Software and UPM – Madrid, ES)

We computer scientists seem to do it differently to other sciences: we publish mostly in conferences and our conferences are of a different nature. Our journal papers are long and take a long time to review and publish whereas often their papers are short and published quickly. And all this interacts with the tendency to evaluate researchers or departments frequently and in a mechanical way (via paper numbers and citation counts) instead of infrequently and deeply (by actually reading papers) and the fact that the current way in which bibliometry is done makes our papers invisible to the world. This position paper offers my viewpoint on what the problems are and why they are important, and also elaborates on some realistic ways forward. In particular, regarding the issue of conferences vs. journals, it proposes the model adopted a few years back by the logic programming community for its main conference (ICLP) and journal (TPLP, Cambridge U. Press). This model is based on the assumption that CS journal papers can be of two types: rapid publication papers (similar to those of other sciences and also close to our conference papers) as well as the longer journal papers that are traditional in CS. Then, the concrete proposal is to, instead of publishing the traditional conference proceedings, have the papers submitted instead to a special issue of an (indexed) journal which is ready online in time for the conference. The traditional conference reviewing process is also improved (following journal standards for rapid publication papers and special issues) to include at least two full rounds of refereeing and a final copy editing step. I argue that this model offers an evolutionary path that solves a good number of the incompatibility problems with other sciences of the current CS models, without giving up the essence of CS conferences. For this reason I believe that this model (or one of the closely related models being proposed) would be a clear path forward, and relatively straightforward for the community to adopt widely.

This viewpoint was further elaborated in the following position paper and slides presented at the dagstuhl meeting:
- http://www.dagstuhl.de/mat/Files/12/12452/12452.HermenegildoManuel.Paper.pdf
- http://www.dagstuhl.de/mat/Files/12/12452/12452.HermenegildoManuel.Slides.pdf

3.9 Computing Research Publishing at Springer – with a Particular Focus on LNCS

Alfred Hofmann (Springer-Verlag – Heidelberg, DE)

We first present some figures about the overall STM publishing market to put computer science research publishing in relation to publishing activities in other scientific disciplines; also, several particular features of the STM market are addressed. After positioning Springer as a computer science publisher in comparison to its main competitors and highlighting several special aspects in computer science research publishing, we present Springer as a leading full-scope publisher in computer science in more detail, with a particular focus on Lecture Notes in Computer Science. We demonstrate that Springer supported computer science research community development and (self -)organization in manifold ways and continues to do so, also by experimenting with novel approaches to research result publishing.

3.10 Open Archives and the Invisible College: driving forces towards a new publication model

Nicolas Holzschuch (Inria Rhône-Alpes, FR)

URL http://hal.inria.fr/hal-00749806

Publication methods have been and are constantly evolving in Computer Science. My position, expressed in this paper, is that in this current landscape, there is room for a new publication model, combining OpenArchives for immediate access with editorial peer-reviewing. A key interest of this proposal is that peer-reviewing happens after the results have been released. I start by reviewing recent changes in publication methods in Computer Science and some of their consequences, drawing practical examples from Computer Graphics. I then review existing forces that are contributing to changes in publication methods, and how these forces push towards a new publicationmodel. I describe this model, and review practical requirements to make it work.

3.11 The Relevance of Journals to Computer Science

Srinivasan Keshav (University of Waterloo, CA)

Every scientific discipline builds on the past: new ideas invariably appear from the analysis, synthesis, and repudiation of prior work. It is necessary that records of prior work be as free from error as humanly possible. However, conference publications, by their very nature, are susceptible to errors. A field that treats conferences as archival publications is building on a foundation of sand. I believe, instead, that we should restore the integrity of archival journal publications by taking steps such as reducing publication delays, increasing the pool of journal reviewers, and removing artificial page limits.

3.12 The best science? The best scientists?

Ursula Martin (Queen Mary University of London, GB)

We draw attention to some effects of the computer science publication culture, looking at ways in which this affects the profile and career trajectory of individuals when compared other scientists, both within and beyond computer science. We highlight in particular the huge waste of effort of authors and referees brought about by overpublication and conferences with low submission rates; the way in which the conference is biased towards those with the resources (money, time, family circumstances, political freedom) to attend and this may miss out on the best science; and the difficulties that can arise for individuals and the discipline when decisions are made by other scientists and administrators do not understand the computer science publication culture.

3.13 Community, Conversation, and Evaluation versus CISE Publication Trends

Keith Marzullo (NSF – Arlington, US)

Scientific journals appeared about 350 years ago to create communities of scientists (at that time, professional societies) and to archive "conversations" among this community. Much more recently, scientific publication has become important as a way to evaluate scientists for promotion and tenure. In computer science and information sciences, and to some degree computer engineering, conference proceedings now dominate scientific journals in the venue of choice for publication. The effects of this transformation have had well-known perverse effects as well as some less-well known ones, but it is hard to see how we can go back. I would like to review the perverse effects and give some ideas of how the forces that have led to this transformation might be used to help alleviate some of the perverse effects.

3.14 Publishing for the Citation Index.– The Subtle Influence of Instant Rating Tools

Friedemann Mattern (ETH Zürich, CH)

The performance of an author can now be evaluated with our fingertips instead of our intellect – thanks to tools like "google scholar" that analyze the publication web and compute bibliometric indicators (such as citation counts) in real-time. Or do these tools just measure popularity? And does this correlate with quality? Anyhow, scientists understand that evaluators, peers, and search committees increasingly rely on such tools. They adapt to that and keep the citation indexes in mind when they publish. This does not go without influence on our publication culture.

3.15 Journals versus Conferences

Kurt Mehlhorn (MPI für Informatik – Saarbrücken, DE)

In the first part of my talk I compare personal experiences as PC chair of ICALP 2012 and editor of JACM, TALG, and CGTA. Despite the fact that the former are more pleasant, I nevertheless believe that CS needs a strong journal culture. As a reader I prefer polished journal articles over conference papers and as an author I like to archive my results in polished form.

In the second part of the talk I describe a very successful (since 10 years) new journal format used in the geophysics community. I combines rapid publication as discussion papers with in-depth reviewing for the journal.

3.16 Towards more constructive reviewing of CS papers

Jeff Mogul (HP Labs – Palo Alto, US)

Many people in CS have expressed concerns about an increasingly "hypercritical" approach to reviewing, which can block or discourage the publication of innovative research. The SIGCOMM Technical Steering Committee (TSC) has been addressing this issue, with the goal of encouraging cultural change without undermining the integrity of peer review. Based on my experience as an author, PC member, TSC member, and occasional PC chair, I examine possible causes for hypercritical reviewing, and offer some advice for PC chairs, reviewers, and authors. My focus is on improving existing CS publication cultures and peer review processes, rather than on proposing radical changes.

3.17 Computer Science publication culture: where to go from here?

M. Tamer Özsu (University of Waterloo, CA)

It is well known that computer science follows a publication policy that is predominantly conference-based. Although this dependence on conferences has helped the field in its initial growth years, it is now starting to have a negative impact on the field. I find it inevitable that we will be moving to a journal-based publication culture. The question is how we start from where we are and end up there. I offer some suggestions for some steps that we can take.

3.18 Books, conferences, open publications: culture and quality

José Palazzo Moreira de Oliveira (UFRGS – Porto Alegre, BR)

My position, exposed in this paper, is that the emphasis on conference publishing, in open publishing and the decline of journal publishing are not a specific Culture in Computing Research but the materialization of the new world supported by the Information Technologies. As the Computing Research Community has the domain of the technology it was possible to integrate this knowledge into our cultural environment. The problem now is to change the minds to accept that quality is not tied with the physical appearance of the media but it is intrinsically associated with the quality of the content. We are living in a changing world and a really great challenge is to support the new reality against the established culture.

3.19 Publication culture in computing research – a perspective from a publisher

Sweitze Roffel (Elsevier Publishing – Amsterdam, NL)

Since the 1960's publishing has undergone many changes, mostly driven by technological developments. But technology only impacts the creation and dissemination of knowledge to a certain extent, and in this talk I'll try to give a publisher's perspective of some main drivers in worldwide media today and try to separate what seems to be happening from what people might perceive to be happening. Using examples from Elsevier research into academia I'll compare other fields with the specific publication culture in computer science and touch upon its conferences, new and traditional publication models, the many different actors and their various business models – and its specific research, reviewing, tenure, technical, infrastructural , and "let's re–name the sub field" culture. Seeing technology also allows us to move from processing digital equivalents of paper to much richer forms of knowledge management, I'll present some early attempts at doing so for the computing sciences, highlighting how the many actors need to learn to cooperate as well as compete in an increasingly distributed environment. Technically, organizationally, and with regard to shared standards and infrastructure. This view from a publisher aims to help the discussion on how we can all contribute to better disseminate and promote the enormous creativity and research contributions of the computing sciences.

3.20 Considerations on open access publication and on the role and publication of conference proceedings

Vladimiro Sassone (University of Southampton, GB)

The Computing community makes a very significant use of conferences as a vehicle for the publication of short papers. Like in many other research fields, computing conferences provide an excellent context for early dissemination of results and interaction within the research community on ongoing research. Differently from many other fields, however, computing conferences –some indeed more than others– carry a highly-valued publications, which arguably absorb a high proportion of the community's overall workload. The question therefore arises as to whether such an effort is well spent, or whether an alternative strategy might be more profitable. One problem is that conference publications are not currently indexed by official collectors of bibliometrics. Most often, later journal papers based on conference publications do not receive a significant number of citations, because authors keep citing the original conference paper. Are we damaging ourselves with respect to other scientific communities by publishing our best results in papers whose bibliometrics do not matter? Should learn to publish conference proceedings in a bibliometrics savvy way? Or would we be better advised to de-emphasise the value of conference publications? It is easy to develop an inflated sense of the wider impact of a conference, and as long as the reward from publication in our best conferences is sufficient for career progression, the incentive is taken away from journal publication. And in fact, several significant results end up to never to be published as full articles. Also, as they become increasingly perceived as a vehicle for publication, conferences becomes a very expensive way to publish and partially lose their interactive "raison d'être". An issue closely related to the above regards open access to research and its copyright status. Whilst the behaviour of professional publishers remains at time questionable, some practical and long-term concerns seem to exist on community-driven publications. How to monitor production quality, guarantee long-term open access availability, defend author-retained copyrights, etc, without overburdening the research community?

3.21 Impact of Publications Culture

Fred B. Schneider (Cornell University – Ithaca, US)

The nature and role of various styles of publication and other vehicles associated with disseminating and/or incentivizing research has changed over the last two decades. (Conferences vs Journals is just one dimension.) And some researchers now voice concern that these changes are not in the best interests of the field, citing anecdotes to justify adopting new mechanisms to create different incentives and presumably reinforce improved values for the field and society.

I submit: before proposing solutions we need to understand the problem. Specifically, we need to understand how our scientific culture is changing and what are the consequences of those changes. To what extent is the nature of research that computer scientists undertake determined by the publications culture and how will changes in the culture affect the research enterprise? What are the costs to the field of various aspects of our current practices and how might those costs be expected to change?

3.22 Publish Now, Judge Later

Douglas B. Terry (Microsoft Research – Mountain View, US)

Conferences these days face a reviewing crisis with too many submissions and not enough time for reviewers to carefully evaluate each submission. Conferences boast about their low acceptance rates as if this were the main metric for evaluating the conference's quality. Numerous good papers get rejected and are resubmitted many times over to different conferences before the paper is eventually accepted or the authors give up in frustration. Good ideas go unpublished or have their publication delayed, to the detriment of the research community. Poor papers get rejected with little attention and do not get the constructive feedback necessary to improve the paper or the work.

My proposed solution is simple: Conferences should accept and publish **all** reasonable submissions. A submission is "reasonable" if it contains something new (a novel idea, new experimental results, validation of previous results, new way of explaining something, etc.), explains the novelty in a clear enough manner for others to learn from it, and puts the new results in a proper context, i.e. compares the results fairly to previous work. The role of reviewers, rather than looking for reasons to reject a paper or spending time ranking papers, is (a) to assess whether the submission is reasonable according to this criteria, and, perhaps more importantly, (b) to offer concrete suggestions for improving the paper. Ultimately, papers will be judged in the fairness of time by their citation counts and impact on the industry. The “10 years after” or “hall of fame” awards should be used as the way to honor the best papers (as well as publishing them in journals), and these awards should be noted in the ACM Digital Library.

3.23 To Boycott or Not to Boycott

Moshe Y. Vardi (Rice University, US)

There has been sound and fury in the Open Access movement over the past sever months. In December 2011, The Research Works Act (RWA) was introduced in the U.S. House of Representatives. The bill contained provisions to prohibit open-access mandates for federally funded research, effectively nullifying the National Institutes of Health’s policy that requires taxpayer-funded research to be freely accessible online. Many scholarly publishers, including the Association of American Publishers (AAP), expressed support for the bill.

The reaction to the bill and its support by scholarly publishers has been one of sheer outrage, with headlines such as “Academic Publishers Have Become the Enemies of Science.” On January 21, 2012, renowned British mathematician Timothy Gowers declared a boycott on Elsevier, a major scholarly publisher, in a detailed blog posting. The boycott movement then took off, with over 12,000 scientists having joined it so far.

Frankly, I do not understand why Elsevier is practically the sole target for the recent wrath directed at scholarly publisher. Elsevier is no worse than most other commercial publishers, just bigger, I believe. Why boycott Elsevier and not Springer, for example?

Beyond the question of whom to target with a boycott, there is the question of the morality of the boycott. Of course, authors can choose the publications of their choice. Also, as a scholar, I can chose which publications I am willing to support by becoming an editor. but the boycott petition also asks signatories to refrain for refereeing articles submitting to Elsevier journals. This means that if you sign this petition than, in effect, you are boycotting your colleagues who have disagreed with you and chose to submit their articles to an Elsevier journal.

I believe an keeping science separate from politics. If it ok to boycott because publishing politics, why is it not ok to boycott for other political considerations? Is it ok to boycott British journals because of objections to the monarchy? Where do you draw the line to avoid politicizing science?

My perspective is that what really propelled the open-access movement was the continuing escalation of the price of scientific publications during the 1990s and 2000s, a period during which technology drove down the cost of scientific publishing. This price escalation has been driven by for-profit publishers. In the distant past, our field had several small- and medium-sized for-profit publishers. There was a sense of informal partnership between the scientific community and these publishers. That was then. Today, there are two large and dominant for-profit publishers in computing. These publishers are thoroughly corporatized. They are businesses with one clear mission.to maximize the return on investments to their owners and shareholders. At the same time, the scientific community, whose goal is to maximize dissemination, continues to behave as if a partnership exists with for-profit publishers, providing them with content and editorial services essentially gratis. This is a highly anomalous arrangement, in my personal opinion. Why should for-profit corporations receive products and labor essentially for free?

Nevertheless, I do not believe that boycott is the solution. Beyond the moral issue that I raised above, there is a major practical issue. For-profit publishers play a key role in computing-research publishing. For example, Spring, through its Lecture Notes in Computer Science series is probably the largest conference-proceedings publisher in computing. It plays an absolutely critical role in the computing-research ecosystem.

If we want to drive the for-profit publishers out of business, we have to do it the old-fashion way, by out-publishing them. If the professional associations in computing research would expand their publishing activities considerably, they should be able to attract the bulk of computing articles. Even if this will not drive the for-profit publishers out of the computing-research publishing business, it would force them to reform their business practices, which is, after all, what we should be after.

3.24 Future Publishing

Andrei Voronkov (University of Manchester, GB)

We discuss future publishing, including open access, licenses and search for publications.

3.25 The USENIX Association: A Financial Case Study for Open Access

Dan Wallach (Rice University, US)

Many of our professional societies rely on the revenues from paid-access publications to support their ongoing activities. The desire for open access must necessarily compete with keeping the books balanced. This short paper discusses the recent financial history of the USENIX Association, a non-profit professional society that manages a number of top-tier academic conferences in computer systems, networking, and security. Every USENIX publication has been freely available online since 2007, making it an interesting case in point. I'll review USENIX's public financial statements and consider how they have managed.

3.26 The Good, the Naïve, and the Ugly – to Index, or not to Index: that is the Question

Reinhard Wilhelm (Universität des Saarlandes, DE)

DBLP, the bibliographic database for Informatics, developed by Michael Ley at Trier University, is currently expanded in a collaborative project between Trier University and the Leibniz Center for Informatics in Schloss Dagstuhl. The coverage of the computing literature is vastly increased since the project started. Although the man power behind DBLP is increased due to the project, capacity is still limited. This forces the DBLP team to set priorities for publication venues wanting to be indexed in DBLP. A set or rules and a process is being developed to take indexing decisions by a board of scientists. I present the background, the problems, and ask the participants for advice on defendable criteria for indexing decisions.

3.27 Where to send your paper?

Jan van Leeuwen (Utrecht University, NL)

The present publication culture in computer science favors publishing in conference proceedings or even in the arXiv over publishing in journals. To restore the balance and regain a leading role in our dynamic field, the journal tradition in computer science should scale up to meet the needs of the modern times. In particular, computer science needs more high-quality journals with a short submission-to-publication time, monthly or even bi-weekly issues (if the notion of issue is to be maintained at all), and open-access. Scattered initiatives are already beginning to change the scene. The 1999 guideline of Patterson et al. [1] needs amending, to guide the assessment of publications and their impact in the Internet Age.

References

1 D. Patterson, L. Snyder, J. Ullman, Evaluating computer scientists and engineers for promotion and tenure, Best Practice Memo, Computing Research Association, September 1999, http://cra.org/resources/bp-view/evaluating_computer_scientists_and_engineers_for_promotion_and_tenure/.

4 Working Groups

During the workshop, a collaborative effort based on the concept of a “Future Workshop” [1] was undertaken to reveal the main problems in the publication culture of computing research and to identify potential solutions to these problems. The “Future Workshop” was suggested and moderated by Batya Friedman. On Wednesday and Thursday, based on the topical sessions, all participants were asked to submit brief statements about shortcomings. These statements were then classified by Batya Friedman with the help of Lance Fortnow, Nicolas

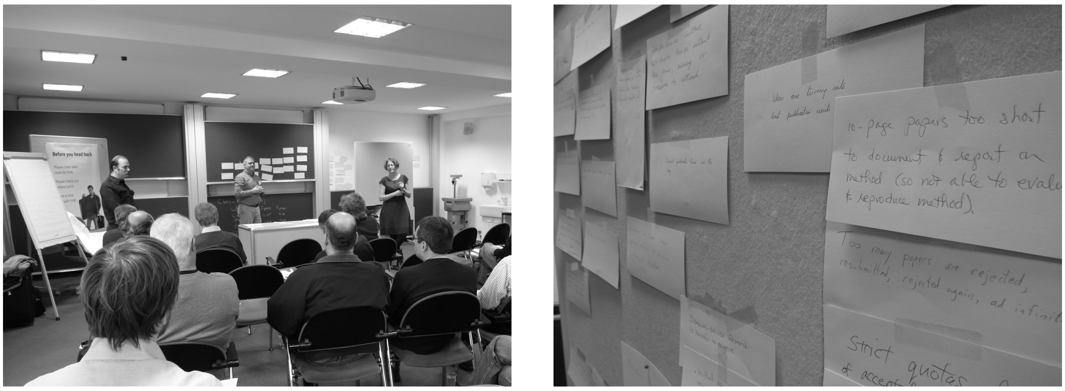

10-page papers too short to document & report on method (so not able to evaluate & reproduce method)
Too many papers are rejected, resubmitted, rejected again, ad infinit
Strict quotas

Figure 1 (left) Batya Friedman moderated the "Future Workshop" sessions. (right) Sample statements submitted by the participants.

Figure 2 The workshop participants discussed and formulated potential remedies for the main problems in the publication culture of computing research.

Holzschuch, and Srinivasan Keshav. Fig. 1 shows a sample of statements as collected during the workshop.

In the second phase of the "Future Workshop" the participants were asked to submit potential solutions. These actions raised a lot of discussions and interactions among the participants and helped to focus on the common grounds of the suggested remedies. See Fig. 2 for some impressions.

In the next phase all participants were asked to "vote" for the most precise and insightful statements. This resulted in a number of statements which in turn were used as basis for working groups. People gathered in front of their favorite statements; leading to working groups of 3-5 participants. The working groups spread out to intensively discuss their chosen topic. As outcome, the working groups provided a brief but concise problem statements along with suggestions how to address these problems. These outcomes are summarized in the following sections.

References

1 F. Kensing, K.H. Madsen, "Generating visions: future workshops and metaphorical design," in J. Greenbaum, M. Kyng, (eds), "Design at work", pp. 155–168, ISBN 0-8058-0612-1, L. Erlbaum Associates Inc., USA, 1992.

4.1 Working Group "On incentives for paper reviewing"

Participants: *Ronald Boisvert, Maarten Fröhlich, Srinivasan Keshav, and Douglas B. Terry*

Our breakout group, consisting of Ron Boisvert, Maarten Fröhlich, Srinivasan Keshav, and Doug Terry, discussed four main issues:

- What incentives exist today for paper reviewing?
- Are there incentives to do good paper reviews?
- What does it mean to do a good paper review?
- What incentives can be put into place to encourage good reviewing?

We all agreed that there are few incentives for reviews to be done in the first place, let alone done well. For conference papers, there are two incentives. First, to gain status by publicly known to be on the TPC. Second, to gain access to relevant papers before publication, privileged information that gives the TPC member an edge [see the reference at the end of this note for more details]. For journal papers, however, there appear to be no incentives at all, other than, perhaps, to curry favour with the journal editor. It was pointed out that in some research areas, such as Mathematical Software, there are no conferences, and TOMS journal is the only publication venue, which puts these research areas at a disadvantage in terms of attracting reviewers.

If someone were to accept to do a review, what incentives exist to do a good review instead of a cursory one? For conferences where reviewers can view each other's reviews, peer pressure guarantees a minimal review quality. Of course, altruistic reviewers, who wish to ensure a high quality conference (or journal, for that matter), go well beyond this base quality level. Later in this abstract, we discuss incentives to do good reviews.

What does it mean to review well in the first place? We think that a good review has the following qualities:

- Timeliness
- Constructive: help improve paper quality rather than finding reasons to reject (hypercriticality)
- Is based on a careful and thorough reading of the paper
- Is unbiased
- Is looking for value, not looking to reject; that is calling out what is good, not just what is bad
- Carefully checks proofs + associated data if applicable
- Validates evaluation results and their statistical significance, if applicable

We also came up with a list of potential incentives for conferences and journals. A simple first step would be to create a "best practices checklist" that could potentially become an ACM standard. This could build on the best practices document already defined by Bertrand Meyer. Another potential approach is to have second-round reviewers (TPC members) rate the first-round reviews. This is already being done in some conferences and serves to increase peer pressure for good reviews. On a more positive note, some conferences are handing out best reviewer awards.

A different tactic is to improve quality by slightly de-anonymizing the reviewers. Public (signed) reviews could be encouraged and the names of the reviewers could be published with accepted papers (as is done with journals today). Potentially, good reviews could earn points, and the reward for a good rating would be that your paper would be reviewed by reviewers also with good ratings, or perhaps you would be asked to be on the TPC/Editorial Board.

A third type of approach would be to maintain a database of reviewer quality assessments by conference chairs and journal editors. The American Physical Society not only does this but also periodically issues a list of top reviewers; these reviewers are sent a formal letter of commendation, with a copy to their supervisor. This ensures that employees value refereeing. Similarly, Springer gives out a Top 20% award to top reviewers. We are not sure if there might be legal drawbacks when doing this, since the database contains inaccessible private information.

Finally, we thought that reviews written by an applicant may be solicited by tenure and promotion committees. This would highly motivate junior faculty.

Incentive structures are studied more formally in [1].

References

1 J. Crowcroft, S. Keshav, and N. McKeown, "Scaling the Academic Publication Process to Internet Scale," Communications of the ACM, Vol. 52, Issue 1, pp. 27–30, January 2009, http://dx.doi.org/10.1145/1435417.1435430.

4.2 Working Group "Hiring based on counting papers"

Participants: *Marc Herbstritt, Kurt Mehlhorn, Jeff Mogul, and M. Tamer Özsu*

Specific issue: CS departments and other employers expect new PhDs to have lots of papers on their CVs.

- How does this differ from other fields?
- How did we get here? Hypothesis: Growing ratio of applicants to openings, desire to have a quick filter on which cvs to take seriously, plus a self-reinforcing cycle of students behaving strategically and increasing the number of papers on their CVs.

Suggestions that the group considered, for academic departments and research labs that are hiring computer scientists:

Suggestion #1: Require CVs (and also papers as published) show clear indications of whether each author was primary, secondary, or supervisory. (Concern: May encourage students to have even longer CVs.)

Suggestion #2: Make it clear that students will be evaluated based on just one paper, chosen by applicant, supported by a research statement to set context. (Concern: Might also perversely encourage a student to publish too many papers, in order to be able to have 1 "best" to choose from.)

Suggestion #3: Require that CVs include (near top) a single, concise paragraph describing the student's deep and major contribution, and without making reference to number of or venue for published papers. (This is to improve to "quick filter" process and to encourage students to have an identifiable major contribution beyond the single paper.)

The working group did not suggest that these are all good suggestions, especially if followed individually. Possibly it would be a good idea to adopt all three of these as a group, to reduce risk of perverse outcomes.

4.3 Working Group "Wrong incentive in publication industry"

Participants: *Bertrand Meyer, Vladimiro Sassone, and Reinhard Wilhelm*

Question was on the wrong incentive in the publication industry.

From the point of view of publishers, indexing companies, and conference organisers this is clearly money. We felt the economic drivers there make it a tough question, and didn't spend much time on it.

Prompted by Andrei Voronkov, we focussed on the incentives for authors and reviewers.

We thought that it would be interesting to form a pool of reviewers rewarded according to the quality of their reviews. The quality would be assessed by the editors, and also by feedback from the authors, e.g. from a rebuttal phase. Reviewers of journal papers should be paid for their work, and the fee might depend on the quality of the reviewer's reviews.

We focus on the idea of open refereeing. The suggestion is that when writing a report referees should make the explicit choice of whether or not to be anonymous. The positive aspects of this are obvious, as it would promote high quality good refereeing. However, it may have the effect of refereeing trying to please powerful authors. etc.

We discussed the idea that there could be a history attached with each reviewer (and each author). There could then be blacklists of authors and reviewers. But this is dangerous ground, because everybody should be allowed to make mistakes without these being attached to them forever.

Finally, we discussed about best practice for journal editors-in-chiefs and conference PC chairs. The community should develop best practice handbooks.

(During the report phase, Marc Herbstritt pointed out that this exists:
http://publicationethics.org/resources/code-conduct.)

5 Further links, immediate outcome, and follow-up discussions

5.1 Workshop-related resources

- Collection of position statements and slides from seminar participants: http://www.dagstuhl.de/mat/index.en.phtml?12452
- Collection of literature and resources on the web related to the workshop topic: http://www.bibsonomy.org/search/12452
- CRA collection on "Scholarly Publications": http://www.cra.org/scholarlypub/

5.2 Immediate outcome

- Srinivasan Keshav, "What is a good quality paper?", checklist draft. http://blizzard.cs.uwaterloo.ca/keshav/mediawiki-1.4.7/index.php/Evaluating_a_research_paper
- Bertrand Meyer, "Conferences: Publication, Communication, Sanction," Blog@CACM, January 10, 2013. http://cacm.acm.org/blogs/blog-cacm/159380-conferences-publication-communication-sanction/fulltext

- Bertrand Meyer, “The Waves of Publication,” Blog@CACM, January 27, 2013. http://cacm.acm.org/blogs/blog-cacm/160071-the-waves-of-publication/fulltext
- Jonathan Grudin, Gloria Mark, and John Riedl, “Conference-Journal Hybrids”, CACM Viewpoint, CACM, Vol. 56, No. 1, pp. 44–49, 2013. http://dx.doi.org/10.1145/2398356.2398371

5.3 Further links

- European Forum for Information and Communication Sciences and Technologies (ICST): http://www.eficst.eu/.
- IEEE black list of conferences: http://www.ieee.org/conferences_events/conferences/publishing/author_form.html.
- An alternative view of scam at IEEE: http://anti-ieee.blogspot.de/2008/02/scam-ieee.html.
- Presentations of Anne-Wil Harzing about scam and fraud in scholarly publishing:
  - How to become a highly cited ESI author: http://www.harzing.com/esi_highcite.htm
  - Predatory “open access” (i.e. author pays) practices: http://www.harzing.com/download/predatoryoa.pdf
- Brian Osserman (UC Davis, US), “Improving the Refereeing Process: A Simple Proposal” (Notices of the AMS, Nov. 2012): http://www.ams.org/notices/201210/rtx121001383p.pdf
- Slow Science Movement:
  - Manifesto: http://slow-science.org/
  - Cross-check by John Horgan in the *Scientific American*: http://blogs.scientificamerican.com/cross-check/2011/07/29/the-slow-science-movement-must-be-crushed/

Participants

- Andrew P. Bernat
Computing Research Association, US
- Ronald F. Boisvert
NIST – Gaithersburg, US
- George Danezis
Microsoft Research UK – Cambridge, GB
- Lance Fortnow
Georgia Inst. of Technology, US
- Batya Friedman
University of Washington, US
- Maarten Fröhlich
IOS Press, NL
- Jonathan Grudin
Microsoft Res. – Redmond, US
- Marc Herbstritt
Schloss Dagstuhl – Saarbrücken/Wadern, DE
- Manuel Hermenegildo
IMDEA Software – Madrid, ES
- Alfred Hofmann
Springer-Verlag – Heidelberg, DE
- Nicolas Holzschuch
Inria Rhône-Alpes, FR
- Srinivasan Keshav
University of Waterloo, CA
- Ursula Martin
Queen Mary University of London, GB
- Keith Marzullo
NSF – Arlington, US
- Friedemann Mattern
ETH Zürich, CH
- Kurt Mehlhorn
MPI für Informatik – Saarbrücken, DE
- Bertrand Meyer
ETH Zürich, CH
- Jeff Mogul
HP Labs – Palo Alto, US
- M. Tamer Özsu
University of Waterloo, CA
- José Palazzo Moreira de Oliveira
UFRGS – Porto Alegre, BR
- Sweitze Roffel
Elsevier Publishing – Amsterdam, NL
- Vladimiro Sassone
University of Southampton, GB
- Fred B. Schneider
Cornell University – Ithaca, US
- Douglas B. Terry
Microsoft Research – Mountain View, US
- Jan van Leeuwen
Utrecht University, NL
- Moshe Y. Vardi
Rice University, US
- Andrei Voronkov
University of Manchester, GB
- Dan Wallach
Rice University, US
- Reinhard Wilhelm
Universität des Saarlandes, DE / Schloss Dagstuhl – Saarbrücken/Wadern, DE

Report from Dagstuhl Seminar 12461

Games and Decisions for Rigorous Systems Engineering

Edited by

Nikolaj Bjørner[1], Krishnendu Chatterjee[2], Laura Kovacs[3], and Rupak M. Majumdar[4]

1 **Microsoft Research – Redmond, US,** nbjorner@microsoft.com
2 **IST Austria – Klosterneuburg, AT,** Krishnendu.Chatterjee@ist.ac.at
3 **TU Wien, AT,** lkovacs@complang.tuwien.ac.at
4 **MPI for Software Systems – Kaiserslautern, DE,** rupak@mpi-sws.org

Abstract

This report documents the program and the outcomes of the Dagstuhl Seminar 12461 “Games and Decisions for Rigorous Systems Engineering”. The seminar brought together researchers working in rigorous software engineering, with a special focus on the interaction between synthesis and automated deduction. This event was the first seminar of this kind and a kickoff of a series of seminars organised on rigorous systems engineering. The theme of the seminar was close in spirit to many events that have been held over the last decades. The talks scheduled during the seminar naturally reflected fundamental research themes of the involved communities.

Seminar 11.–16. November, 2012 – www.dagstuhl.de/12461

1998 ACM Subject Classification C.1.4 Parallel Architectures, D.2.4 Software/Program Verification, D.2.5 Testing and Debugging, D.3.1 Formal Definitions and Theory, D.3.2 Language Classifications, F.1.1 Models of Computation, F.4.1 Mathematical Logic, I.2.2 Automatic Programming, I.2.3 Deduction and Theorem Proving

Keywords and phrases Systems Engineering, Software Verification, Reactive Synthesis, Automated Deduction

Digital Object Identifier 10.4230/DagRep.2.11.45

1 Executive Summary

Nikolaj Bjørner
Krishnendu Chatterjee
Laura Kovacs
Rupak M. Majumdar

Principled approaches to systems design offer several advantages, including developing safety-critical systems and scaling technological advances with multi-core processes and cloud computing. Rigorous mathematical techniques, such as model checking, decision procedures, and abstract interpretation, are dominantly used a posteriori in systems engineering: a program is formally analyzed after it has been developed. In the context of rigorous systems engineering, post-hoc verification is however very costly and error-prone. The explosion of concurrent computation in the new generation of embedded systems has therefore motivated the integration of established methods with novel techniques in the design process from day one.

Games and Decisions for Rigorous Systems Engineering, *Dagstuhl Reports*, Vol. 2, Issue 11, pp. 45–65
Editors: Nikolaj Bjørner, Krishnendu Chatterjee, Laura Kovacs, and Rupak M. Majumdar
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Such an integration has been materialized in using game theoretic synthesis of reactive systems from higher level design requirements. In many synthesis algorithms, it is better to work with symbolic representations, where the state space is modeled using logical formulas. This enables techniques to scale to potentially infinite models, but requires decision procedures for checking the validity of sentences in the pertinent logical theories. The increasingly complex integration of model checking with complementary techniques such as software testing has imposed new requirements on decision procedures, such as proof generation, unsatisfiable core extraction, and interpolation.

The main goal of the Dagstuhl Seminar 12461 "Games and Decisions for Rigorous Systems Engineering" was to bring together researchers working in the field of rigorous systems engineering, the tool-supported application of mathematical reasoning principles to the design and verification of complex software and hardware systems. The seminar had a special focus on developing systems (reactive, concurrent, distributed) using recent advances in game theoretic synthesis and in decision procedures and automated deduction techniques.

The seminar covered the following three main areas:

- software verification (reactive, concurrent, distributed);
- game theory and reactive synthesis;
- decision procedures (SAT, SMT, QBF) and theorem proving (first and higher order).

Within the scope of these areas, the seminar addressed tooling around software testing, model checking, interpolation, decision procedures, and model finding methods in automated theorem proving.

In the spirit of advancing tools and theory in related areas of theorem proving and model checking, the seminar schedule included tutorials on games, synthesis, theorem proving; research talks on recent results; and discussion sessions on applications and exchange formats for benchmarking tools.

The seminar fell on 5 days in the week of November 12–16, 2012. All together, 43 researchers participated (11 women and 32 men).

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 Lazy Abstraction with Interpolants for Arrays

Francesco Alberti (University of Lugano, CH)

Joint work of Alberti, Francesco; Bruttomesso, Roberto; Ghilardi, Silvio; Ranise, Silvio; Sharygina, Natasha
Main reference F. Alberti, R. Bruttomesso, S. Ghilardi, S. Ranise, N. Sharygina, "Lazy Abstraction with Interpolants for Arrays," in Proc. of 18th Int'l Conf. on Logic for Programming, Artificial Intelligence, and Reasoning (LPAR'12), LNCS, Vol. 7180, pp. 46–61, 2012.
URL http://dx.doi.org/10.1007/978-3-642-28717-6_7

Efficient and automatic model checking of software with unbounded data structure is a long standing scientific challenge. Abstraction/refinement techniques need to be carefully adapted when unbounded data structures come into play because of the need of quantified predicates. In this talk we will describe a recently proposed framework, "Lazy Abstraction with Interpolants for Arrays", suited for reasoning about programs with unbounded arrays. The framework integrates a symbolic backward reachability analysis with an interpolation-based refinement procedure. It allows for an efficient handling of quantified formulas representing backward reachables states and exploiting of quantifier-free interpolation algorithms for refining predicates along spurious counterexamples. The talk will also describe SAFARI (SMT-based Abstraction For Arrays with Interpolants), a tool implementing this framework, and present "term abstraction", a heuristic used to tune interpolation algorithms.

3.2 Conditional Model Checking: A Technique to Pass Information between Verifiers

Dirk Beyer (Universität Passau, DE)

Joint work of Beyer, Dirk; Henzinger, Thomas A.; Keremoglu, M. Erkan; Wendler, Philipp
Main reference D. Beyer, T.A. Henzinger, M.E. Keremoglu, P. Wendler, "Conditional Model Checking: A Technique to Pass Information between Verifiers," in Proc. of the 20th ACM SIGSOFT Int'l Symp. on the Foundations of Software Engineering (FSE'12), ACM, 2012.
URL http://dx.doi.org/10.1145/2393596.2393664
URL http://www.sosy-lab.org/ dbeyer/Publications/2012-FSE.Conditional_Model_Checking.pdf

Software model checking, as an undecidable problem, has three possible outcomes: (1) the program satisfies the specification, (2) the program does not satisfy the specification, and (3) the model checker fails. The third outcome usually manifests itself in a space-out, time-out, or one component of the verification tool giving up; in all of these failing cases, significant computation is performed by the verification tool before the failure, but no result is reported. We propose to reformulate the model-checking problem as follows, in order to have the verification tool report a summary of the performed work even in case of failure: given a program and a specification, the model checker returns a condition P —usually a state predicate— such that the program satisfies the specification under the condition P —that is, as long as the program does not leave the states in which P is satisfied. In our experiments, we investigated as one major application of conditional model checking the sequential combination of model checkers with information passing. We give the condition that one model checker produces, as input to a second conditional model checker, such that the verification problem for the second is restricted to the part of the state space that is not covered by the condition, i.e., the second model checker works on the problems that the first model checker could not solve. Our experiments demonstrate that repeated application of conditional model checkers, passing information from one model checker to the next, can significantly improve the verification results and performance, i.e., we can now verify programs that we could not verify before.

3.3 Efficient Controller Synthesis for Consumption Games with Multiple Resource Types

Tomas Brazdil (Masaryk University, CZ)

Joint work of Brazdil, Tomas; Chatterjee, Krishnendu; Kucera, Antonin: Novotny, Petr

We introduce consumption games, a model for discrete interactive system with multiple resources that are consumed or reloaded independently. More precisely, a consumption game is a finite-state graph where each transition is labeled by a vector of resource updates, where every update is a non-positive number or omega. The omega updates model the reloading of a given resource. Each vertex belongs either to player Box or player Diamond, where the aim of player Box is to play so that the resources are never exhausted. We consider several natural algorithmic problems about consumption games, and show that although these problems are computationally hard in general, they are solvable in polynomial time for every fixed number of resource types (i.e., the dimension of the update vectors) and bounded resource updates.

3.4 Stochastic Program Synthesis with Smoothed Numerical Search

Swarat Chaudhuri (Rice University, US)

Writing programs that behave optimally on probabilistic inputs is a highly challenging task. In this talk, I will describe a program synthesis procedure targeted to such tasks. The procedure takes an input an infinite-state program sketch annotated with a set of boolean assertions and a set of quantitative objectives. The procedure's goal is to find an implementation that (a) satisfies the boolean assertions with probability above a certain bound; (b) in the expected behavior of the program, the quantitative objectives are minimized. We solve this problem using a combination of local numerical optimization and probabilistic abstract interpretation, called "smoothed numerical search". The core idea of the algorithm is to approximate a program using a series of smooth mathematical functions. Each of these approximations is "unsound"; however, at the limit they converge to a sound abstraction.

3.5 Games in System Design: Tutorial and Survey

Laurent Doyen (CNRS, ENS – Cachan, FR)

We present a tutorial introduction on two-player games played on graphs, we discuss their fundamental properties, and basic ingredients for algorithmic solutions and complexity analysis. Along with applications in the design and formal verification of reactive systems, we survey recent results about games with quantitative objectives, combination of multiple objectives, and partial-observation games.

3.6 A new learning scheme for QDPLL solvers

Uwe Egly (TU Wien, AT)

Joint work of Egly, Uwe; Lonsing, Florian; Van Gelder, Allen

Most of todays DPLL-based QBF solvers employ Q-resolution to learn clauses or cubes. The classical learning scheme from conflicts starts with the clause falsified by the current assignment and resolve upon existential variables in reverse assignment order using antecedent clauses stored in the implication graph during BCP.

We identify a class of formulas F1, F2, ... for which (i) learning a single clause with the above scheme in an evaluation of Fk is exponential in k and (ii) Fk has a resolution proof of linear length. We propose a new learning scheme which avoids such efficency problems. The new scheme employs resolution “from the decisions towards the conflict”, i.e., in the direction in which decisions have been propagated during QBCP. We discuss experimental results obtained with a very first implementation into depQBF.

This is joint work with Florian Lonsing (Vienna University of Technology) and Allen Van Gelder (University of California at Santa Cruz).

3.7 Inductive Data Flow Graphs

Azadeh Farzan (University of Toronto, CA)

Joint work of Farzan, Azadeh; Kincaid, Zachary; Podelski, Andreas;
Main reference To appear in POPL 2013 (proceedings information not available yet).

The correctness of a sequential program can be shown by the annotation of its control flow graph with inductive assertions. We propose inductive data flow graphs, data flow graphs with incorporated inductive assertions, as the basis of an approach to verifying concurrent programs. An inductive data flow graph accounts for a set of dependencies between program actions in interleaved thread executions, and therefore stands as a representation for the set of concurrent program traces which give rise to these dependencies. The approach first constructs an inductive data flow graph and then checks whether all program traces are represented. The size of the inductive data flow graph is polynomial in the number of data dependencies (in a sense that can be made formal); it does not grow exponentially in the number of threads unless the data dependencies do. The approach shifts the burden of the exponential explosion towards the check whether all program traces are represented, i.e., to a combinatorial problem (over finite graphs).

3.8 Synthesis of reactive systems

Bernd Finkbeiner (Universität des Saarlandes, DE)

Joint work of Finkbeiner, Bernd; Schewe, Sven; Jacobs, Swen

More than fifty years after its introduction by Alonzo Church, the synthesis problem is still one of the most intriguing challenges in the theory of reactive systems. Synthesis is particularly difficult in the setting of distributed systems, where we try to find a combination of process implementations that jointly guarantee that a given specification is satisfied. A reduction from multi-player games shows that the problem is in general undecidable. Despite this negative result, there is a line of discoveries where the decidability of the synthesis problem was established for distributed systems with specific architectures, such as pipelines and rings, or other restrictions on the problem, such as local specifications. Encouraged by these findings, new specification languages like Coordination Logic aim for a comprehensive logical representation and a uniform algorithmic treatment of the decidable synthesis problems. In this talk, I will trace the progress from isolated decidability results towards universal synthesis logics and algorithms. I will demonstrate how the logical representation of the synthesis problem simplifies the identification of decidable cases and give an overview on the state of the art in decision procedures and strategy construction algorithms for the synthesis of reactive systems.

3.9 Deciding Floating-Point Logic with Systematic Abstraction

Alberto Griggio (Fondazione Bruno Kessler – Trento, IT)

Joint work of Haller, Leopold; Griggio, Alberto; Brain, Martin; Kroening, Daniel;
Main reference L. Haller, A. Griggio, M. Brain, D. Kroening, "Deciding Floating-Point Logic with Systematic Abstraction," in Proc. of the 12th Conf. on Formal Methods in Computer-Aided Design (FMCAD'12), 2012.
URL
URL http://www.cs.utexas.edu/ hunt/FMCAD/FMCAD12/fmcad2012.pdf

We present a bit-precise decision procedure for the theory of binary floating-point arithmetic. The core of our approach is a non-trivial generalisation of the conflict analysis algorithm used in modern SAT solvers to lattice-based abstractions. Existing complete solvers for floating-point arithmetic employ bit-vector encodings. Propositional solvers based on the Conflict Driven Clause Learning (CDCL) algorithm are then used as a back-end. We present a natural-domain SMT approach that lifts the CDCL framework to operate directly over abstractions of floating-point values. We have instantiated our method inside MathSAT with the floating-point interval abstraction. The result is a sound and complete procedure for floating-point arithmetic that outperforms the state-of-the-art significantly on problems that check ranges on numerical variables. Our technique is independent of the specific abstraction and can be applied to problems beyond floating-point satisfiability checking.

3.10 Concurrent Test Generation using Concolic Multi-Trace Analysis

Aarti Gupta (NEC Laboratories America, Inc. – Princeton, US)

Joint work of Gupta, Aarti; Razavi, Noloofar; Kahlon, Vineet

Discovering concurrency bugs is inherently hard due to nondeterminism in multi-thread scheduling. Predictive analysis techniques have been used to find such bugs by observing given test runs, and then searching for other interesting thread interleavings. For sequential code, SMT-based concolic execution techniques have been used successfully to generate interesting test inputs to increase structural code coverage such as branch or statement coverage. In this talk, I will describe our recent work that targets increasing code coverage in multi-thread programs by using a concolic multi-trace analysis (CMTA) that combines elements of prediction with generation of new test inputs. We have implemented CMTA and show encouraging results on benchmark programs.

This is joint work with Niloofar Razavi, Franjo Ivancic, and Vineet Kahlon; to appear soon at the Asian Symposium on Programming Languages and Systems (APLAS 2012).

3.11 VINTA: Verification with Interpolation and Abstract Interpretation

Arie Gurfinkel (CMU – Pittsburgh, US)

Joint work of Gurfinkel, Arie; Albarghouthi, Aws; Chechik, Marsha
Main reference A. Gurfinkel, A. Albarghouthi, M. Chechik, "Craig Interpretation," in Proc. of 19th Int'l Symp. on Static Analysis (SAS'12), LNCS, Vol. 7460, pp. 300–316, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-33125-1_21

Abstract interpretation (AI) is one of the most scalable automated program verification techniques. The scalability is achieved through aggressive abstraction in basic analysis steps (i.e., joins and widening). This leads to loss of precision. As such, AI is plagued by false alarms. In this talk, I will present VINTA, an algorithm that enriches AI with Abstraction Refinement techniques from Model Checking to alleviate the false alarms. VINTA is an iterative algorithm that uses Craig interpolants to refine and guide AI away from false alarms. VINTA is based on a novel refinement strategy that capitalizes on recent advances in SMT and interpolation-based Model Checking. On one hand, it can find concrete counterexamples to justify alarms produced by AI. On the other, it can strengthen invariants to exclude alarms that cannot be justified. The refinement process continues until either a safe inductive invariant is computed, a counterexample is found, or resources are exhausted. This strategy allows VINTA to recover precision lost in many AI steps. VINTA has been implemented as part of the UFO verification framework. It is a big contributor to the success of UFO in the 2nd International Software Verification Competition.

3.12 Proof Tree Preserving Interpolation

Jochen Hoenicke (Universität Freiburg, DE)

Craig interpolants are widely used in model checking and state space abstraction. Interpolants typically are extracted from proofs produced by theorem provers. While this extraction procedure is easy and well understood in the context of propositional logic, extracting interpolants from a proof generated by an SMT solver is more complex. In contrast to SAT solvers, SMT solvers create new literals, e.g., to combine multiple theories in a Nelson-Oppen style or to split the solution space using cuts. These literals might contain symbols local to different parts of the interpolation problem. Such literals are called mixed, or, sometimes, uncolorable. Resolution steps on mixed literals are the major difficulty when extracting interpolants from proofs from SMT solvers.

We present a technique to compute Craig interpolants in the theory of uninterpreted functions combined with the theory of linear arithmetic either over the reals or the integers. The interpolation scheme is based on a syntactical restriction of the partial interpolants and specialized rules to interpolate resolution steps on mixed literals. Contrary to existing approaches, this scheme neither limits the inferences done by the SMT solver, nor does it transform the proof tree before extracting interpolants. The interpolation scheme is used in the interpolating SMT solver SMTInterpol.

3.13 Underapproximation of Procedure Summaries for Integer Programs

Radu Iosif (VERIMAG – Gières, FR)

We show how to underapproximate the procedure summaries of recursive programs over the integers using off-the-shelf analyzers for non-recursive programs. The novelty of our approach is that the non-recursive program we compute may capture unboundedly many behaviors of the original recursive program for which stack usage cannot be bounded. Moreover, we identify a class of recursive programs on which our method terminates and returns the precise summary relations without underapproximation. Doing so, we generalize a similar result for non-recursive programs to the recursive case. Finally, we present experimental results of an implementation of our method applied on a number of examples.

3.14 Proving Properties about Functional Programs

Moa Johansson (Chalmers UT – Göteborg, SE)

Joint work of Claessen, Koen; Johansson, Moa; Rosen, Dan; Smallbone, Nicholas

HipSpec is an automatic inductive theorem prover for proving properties about Haskell programs. It implements a novel bottom-up approach to lemma discovery: potentially interesting lemmas about available functions and datatypes are first synthesised creating a richer background theory for the prover.

HipSpec consists of several sub-systems: Hip is an inductive theorem prover. It translates Haskell function definitions to first order logic and applies induction to given conjectures. Resulting proof obligations are passed to an off the shelf prover (for instance E or Z3).

QuickSpec is responsible for generating candidate lemmas about available functions and datatypes. It generates terms which are divided up into equivalence classes using counterexample testing. From these equivalence classes, equations can be derived. These are passed to Hip for proof. Those that are proved are added to the background theory and may be used in subsequent proofs.

HipSpec is available for download from https://github.com/danr/hipspec

3.15 Preprocessing for first-order logic with applications to hardware verification

Konstantin Korovin (University of Manchester, GB)

Joint work of Krystof Hoder, Zurab Khasidashvili, Konstantin Korovin and Andrei Voronkov
Main reference K. Hoder, Z. Khasidashvili, K. Korovin, A. Voronkov, "Preprocessing Techniques for First-Order Clausification," in Proc. of the 12th Conf. on Formal Methods in Computer-Aided Design (FMCAD'12), pp. 44–51, 2012.
URL http://www.cs.man.ac.uk/ korovink/my_pub/fmcad_2012.pdf

We discuss several preprocessing techniques for simplifying first-order formulas aimed at improving clausification. These include definition inlining and merging, simplifications based on new data structure called quantified AIG, and its combination with OBDDs. These techniques were inspired by applications of first-order theorem proving to hardware verification.

3.16 Asynchronous Games over Tree Architectures

Anca Muscholl (Université Bordeaux, FR)

The control problem starts with a plant and asks to restrict its controllable actions in such a way that a given specification is met. Synthesis can be seen as a special case of control. We consider a distributed version of the control problem where both plant and controller are parallel compositions of finite-state processes communicating via shared variables (also known as Zielonka asynchronous automata). The most important aspect of this model is that the processes participating in a synchronization can exchange the complete information about their causal past. Thus, it is an intriguing open problem whether this control setting is decidable. We show decidability when the communication architecture is acyclic. In this case, if there is a solution then controllers exchange only bounded information. The complexity of our algorithm is l-fold exponential with l being the height of the tree representing the architecture. We show that this complexity is tight.

3.17 A Tutorial on SAT and SMT

Albert Oliveras (TU of Catalonia – Barcelona, ES)

In this tutorial, an overview of SAT and SMT will be given.

In the first part, we will introduce the problem of SAT and its state-of-the-art techniques. Starting from an abstract presentation of the DPLL procedure, we then explain how several conceptual enhancements can be added to it giving the so-called CDCL algorithm for SAT solving.

After that, we will focus on the problem of SMT. We first overview the most common theories that SMT solvers deal with. Then, we focus on the two main approaches to SMT: the eager and the lazy approach, making special emphasis on the last one. In particular, we make clear which are the requirements that a theory solver needs to have to be used in a DPLL(T) system.

3.18 Decision Problems for Linear Recurrence Sequences

Joel Ouaknine (University of Oxford, GB)

Joint work of Ouaknine, Joel; Worrel, James; Daws, Matt

Linear recurrence sequences (such as the Fibonacci numbers) permeate a vast number of areas of mathematics and computer science, and also have many applications in other fields such as economics and theoretical biology. In the context of synthesis and verification, linear recurrence sequences arise in connection with linear programs, probabilistic systems, stochastic logics, and linear dynamical systems, among others.

In this talk, I will focus on three fundamental decision problems for linear recurrence sequences, namely the Skolem Problem (does the sequence have a zero?), the Positivity Problem (is the sequence always positive?), and the Ultimate Positivity Problem (is the sequence ultimately always positive?).

3.19 Automated Game-theoretic Verification for Probabilistic Systems

David Parker (University of Birmingham, GB)

Joint work of Chen, Taolue; Forejt, Vojtěch; Kwiatkowska, Marta; Parker, David; Simaitis, Aistis
Main reference T. Chen, V. Forejt, M. Kwiatkowska, D. Parker, A. Simaitis, "Automatic Verification of Competitive Stochastic Systems," in Proc. of the 18th Int'l Conf. on Tools and Algorithms for the Construction and Analysis of Systems (TACAS'12), LNCS, Vol. 7214, pp. 315–330, Springer, 2012
URL http://dx.doi.org/10.1007/978-3-642-28756-5_22

We present automatic verification techniques for turn-based stochastic multi-player games, which model probabilistic systems containing components that can either collaborate or compete in order to achieve particular goals. We give model checking algorithms for a temporal logic called rPATL, which allows us to reason about the collective ability of a set of players to achieve a goal relating to the probability of an event's occurrence or the expected amount of cost/reward accumulated. We implement our techniques in an extension of the PRISM model checker and use them to analyse and detect potential weaknesses in systems such as algorithms for energy management and collective decision making for autonomous systems.

3.20 Beluga: Programming proofs in context

Brigitte Pientka (McGill University – Montreal, CA)

We routinely reason about the runtime behavior of software using formal systems such as type systems or logics for access control or information flow to establish safety and liveness properties. In this talk, I will give an overview of Beluga, a dependently typed programming and proof environment. It supports specifying formal systems in the logical framework LF and directly supports common and tricky routines dealing with variables, such as capture-avoiding substitution and renaming. Moreover, Beluga allows embedding LF objects together with their context in programs and types supporting inductive and coinductive definitions and we can manipulate contextual LF objects via pattern matching. Taken together these features lead to a powerful language which supports writing compact and elegant proofs.

3.21 Proving termination of C-like programs using MAX-SMT

Albert Rubio (UPC – Barcelona, ES)

We show how MAX-SMT can be used for proving and disproving termination of C-like programs. MAX-SMT allow us to characterize our termination problem giving different weights to the needed conditions, providing a better notion of progress. This also makes it easier to combine the process of building the termination argument with the usually necessary process of generating invariants.

Our technique focuses on proving termination, but sometimes from the already generated invariants and partial termination arguments, we can prove non-termination or warn about potential cases of non-termination.

The method has been implemented in a prototype that has successfully been tested on a wide sample of examples.

3.22 Program verification as constraint solving (also for CTL* properties)

Andrey Rybalchenko (TU München, DE)

Joint work of Beyene, Tewodros; Popeea, Corneliu; Rybalchenko, Andrey;

First, we review how proving reachability and termination properties of transition systems, procedural programs, multi-threaded programs, and higher- order functional programs can be reduced to constraint solving. Second, we show how CTL* properties can be proved using contraint-based setting. Finally, we discuss adequate solving algorithms and tools.

3.23 QBFs and Certificates

Martina Seidl (University of Linz, AT)

Joint work of Seidl, Martina; Niemetz, Aina; Preiner, Mathias; Lonsing, Florian; Biere, Armin
Main reference A. Niemetz, M. Preiner, F. Lonsing, M. Seidl, A. Biere, "Resolution-Based Certificate Extraction for QBF," in Proc. of the 15th Int'l Conf. on Theory and Applications of Satisfiability Testing (SAT'12), LNCS, Vol. 7317, pp. 430–435, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-31612-8_33

A certificate of (un)satisfiability for a quantified Boolean formula (QBF) represents concrete assignments to the variables of the formula. Certificates are not only witnesses for the truth value returned by a QBF solver, but also represent the solutions for practical applications of QBF like formal verification and model checking. Recently, an approach has been presented, which can be directly built on top of DPLL based QBF solvers. Starting from resolution proofs produced by the solver during clause and cube learning, the certificates are constructed by certain syntactic properties of the proof tree. Based on our integrated set of tools realizing resolution-based certificate extraction for QBFs in prenex conjunctive normal form, in this talk, we discuss the state-of-the-art of QBF certification and point out future challenges.

3.24 Incremental Upgrade Checking by Means of Interpolation-based Function Summaries

Natasha Sharygina (University of Lugano, CH)

Joint work of Ondrej, Sery; Fedyukovich, Grigory; Sharygina, Natasha
Main reference S. Ondrej, G. Fedyukovich, N. Sharygina, "Incremental Upgrade Checking by Means of Interpolation-based Function Summaries," in Proc. of the 12th Conf. on Formal Methods in Computer-Aided Design (FMCAD'12), 2012.
URL http://www.verify.inf.unisi.ch/files/fmcad2012.pdf

During its evolution, a typical software/hardware design undergoes a myriad of small changes. However, it is extremely costly to verify each new version from scratch. As a remedy to this problem, we propose to use function summaries to enable incremental verification of the evolving systems. During the evolution, our approach maintains function summaries derived using Craig's interpolation. For each new version, these summaries are used to perform a local incremental check. Benefit of this approach is that the cost of the check depends on the extent of the change between the two versions and can be performed cheaply for incremental changes without resorting to re-verification of the entire system. Our implementation and experimentation in the context of the bounded model checking for C confirms that incremental changes can be verified efficiently for different classes of industrial programs.

3.25 Quantitatively Relaxed Concurrent Data Structures

Ana Sokolova (Universität Salzburg, AT)

Joint work of Henzinger, Thomas A.; Kirsch, Christoph M.; Payer, Hannes; Sezgin, Ali; Sokolova, Ana
Main reference T.A. Henzinger, C.M. Kirsch, H. Payer, A. Sezgin, A. Sokolova, "Quantitative Relaxations of Concurrent Data Structures," in Proc. POPL 2013, to appear.

This talk is about our recent work on relaxing the semantics of concurrent data structures, in a quantitative way, so that they allow better-performing implementations. By their nature, data structures are bottlenecks in the presence of concurrency, e.g. the top pointer of a stack is a point of contention for which all threads compete. As a consequence, implementations of concurrent data structures often show negative scalability. Recent trends in concurrency show that relaxing the semantics may be the way to better performance. In this work we provide a framework for quantitative relaxations of concurrent data structures, where the allowed "error" from the perfect semantics is quantified by a distance. During the talk, we will use a stack as a running example. We also present a new concurrent implementation of a quantitatively relaxed stack that performs well and shows positive scalability.

3.26 Introduction to the Sketch Synthesis System

Armando Solar-Lezama (MIT – Cambridge, US)

I will provide a brief introduction to the Sketch language and highlight some of the recent applications and open problems both in terms of language design and decision procedures.

3.27 A Semantic Account for Modularity in Multi-language Modelling of Search Problems

Eugenia Ternovska (Simon Fraser University – Burnaby, CA)

Joint work of Ternovska, Eugenia; Tasharrofi, Shahab; Wu, Xiongnan

With the increased applications of distributed communicating systems, there is a strong need in formalisms that support modularity. In such systems, the representation language of a module may not even be known outside of that module. I will describe a semantic approach to formal modelling of such systems, and analyze the expressive power of adding a loop operator. I will then describe an algorithmic schema for synthesizing solutions of modular systems. The solutions agree with each of the interacting, collaborating, mutually dependent modules. The algorithmic schema is instantiated with oracle procedures specific to each module. It generalizes ideas underlying "combined" solving such as the DPLL(T) procedure, branch-and-cut ILP solvers and state-of-the-art combination of ASP and CP. Joint work with Shahab Tasharrofi and Xiongnan Wu.

3.28 Secure Two-Party Computation in ANSI C

Helmut Veith (TU Wien, AT)

Joint work of Andreas Holzer, Martin Franz, Stefan Katzenbeisser, and Helmut Veith
Main reference A. Holzer, M. Franz, S. Katzenbeisser, H. Veith, "Secure two-party computations in ANSI C," in Proc. of the 2012 ACM Conf. on Computer and Communications Security (CCS'12), pp. 772–783, ACM, 2012.
URL http://dx.doi.org/10.1145/2382196.2382278
URL http://www.sosy-lab.org/~dbeyer/Publications/2012-FSE.Conditional_Model_Checking.pdf

The practical application of Secure Two-Party Computation is hindered by the difficulty to implement secure computation protocols. While recent work has proposed very simple programming languages which can be used to specify secure computations, it is still difficult for practitioners to use them, and cumbersome to translate existing source code into this format. Similarly, the manual construction of two-party computation protocols, in particular ones based on the approach of garbled circuits, is labor intensive and error-prone.

The central contribution of the current paper is a tool which achieves Secure Two-Party Computation for ANSI C. Our work is based on a combination of model checking techniques and two-party computation based on garbled circuits. Our key insight is a nonstandard use of the bit-precise model checker CBMC which enables us to translate C programs into equivalent Boolean circuits. To this end, we modify the standard CBMC translation from programs into Boolean formulas whose variables correspond to the memory bits manipulated by the program. As CBMC attempts to minimize the size of the formulas, the circuits obtained by our tool chain are also size efficient; to improve the efficiency of the garbled circuit evaluation, we perform optimizations on the circuits. Experimental results with the new tool CBMC-GC demonstrate the practical usefulness of our approach.

3.29 Verification of Low Level List Manipulation

Tomas Vojnar (Brno University of Technology, CZ)

In the talk, we present an ongoing work related to the tool called Predator for verification of programs containing low level list manipulation. We first present some typical problems that arise in low-level list manipulating programs used in system software. Then, we briefly explain how these issues are tackled in Predator using a graph-based representation of sets of heaps (partially inspired by works on separation logic with higher order list predicates, but purely graph-based and significantly extended to cope with the low-level memory manipulation features).

3.30 First-order theorem proving and Vampire

Andrei Voronkov (University of Manchester, GB)

Joint work of Krystof Hoder, Laura Kovács and Andrei Voronkov

In this tutorial we give a short introduction in first-order theorem proving and the use of the theorem prover Vampire.

I will discuss the the resolution and superposition calculus, introduce the saturation principle, present various algorithms implementing redundancy elimination, preprocessing and clause form transformation and demonstrate how these concepts are implemented in Vampire.

I will next also cover more advanced topics and features. Some of these features are implemented only in Vampire. This includes reasoning with theories, such as arithmetic, answering queries to very large knowledge bases, interpolation, and an original technique of symbol elimination, which allows one to automatically discover invariants in programs with loops.

3.31 Labelled Interpolation Systems

Georg Weissenbacher (TU Wien, AT)

Main reference G. Weissenbacher, “Interpolant Strength Revisited,” in Proc. of the 15th Conf. on Theory and Applications of Satisfiability Testing (SAT’12), LNCS, Vol. 7317, pp. 312–326, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-31612-8_24

Craig’s interpolation theorem has numerous applications in model checking, automated reasoning, and synthesis. The intrinsic properties of interpolants enable concise abstractions in verification and smaller circuit designs in synthesis. There is a variety of interpolation systems which derive interpolants from refutation proofs; these systems are ad-hoc and rigid in the sense that they provide exactly one interpolant for a given proof. In this talk, I will discuss how refutation-based interpolation techniques can be parametrised to remove this limitation. Labelled interpolation systems allow for the systematic variation of the logical strength and the size of Craig interpolants. In addition, they generalise a number of existing interpolation techniques for propositional and first-order logic. It is still an open question how applications can exploit the additional flexibility provided by this novel interpolation system.

3.32 Parameterized Model Checking of Fault-tolerant Distributed Algorithms

Josef Widder (TU Wien, AT)

Joint work of John, Annu; Konnov, Igor; Schmid, Ulrich; Veith, Helmut; Widder, Josef
Main reference A. John, I. Konnov, U. Schmid, H. Veith, J. Widder, "Counter Attack on Byzantine Generals: Parameterized Model Checking of Fault-tolerant Distributed Algorithms," arXiv:1210.3846v2 [cs.LO].
URL http://arxiv.org/abs/1210.3846

We introduce a method for automated parameterized verification of fault-tolerant distributed algorithms. The distributed algorithms we consider are parameterized by both the number of processes and the assumed maximum number of Byzantine faulty processes. At the center of our technique is a parametric interval abstraction (PIA) where the interval boundaries are arithmetic expressions over parameters. Using PIA for both data abstraction and a new form of counter abstraction, we reduce the parameterized problem to finite-state model checking. We demonstrate the practical feasibility of our method by verifying several variants of the well-known distributed algorithm by Srikanth and Toueg. To the best of our knowledge, this is the first paper to achieve parameterized automated verification of Byzantine fault-tolerant distributed algorithms.

3.33 Complete Instantiation-Based Interpolation

Thomas Wies (New York University, US)

Joint work of Wies, Thomas; Totla, Nishant

Craig interpolation has been a valuable tool for formal methods with interesting applications in program analysis and verification. Modern SMT solvers implement interpolation procedures for the theories that are most commonly used in these applications. However, many application-specific theories remain unsupported, which limits the class of problems to which interpolation-based techniques apply. In this talk, I present a generic framework to build new interpolation procedures via reduction to existing interpolation procedures. We consider the case where an application-specific theory can be formalized as an extension of a base theory with additional symbols and axioms. Our technique uses finite instantiation of the extension axioms to reduce an interpolation problem in the theory extension to one in the base theory. We identify a model-theoretic criterion that allows us to detect the cases where our technique is complete. We discuss specific theories that are relevant in program verification and that satisfy this criterion. In particular, we obtain complete interpolation procedures for theories of arrays and linked lists. The latter is the first complete interpolation procedure for a theory that supports reasoning about complex shape properties of heap-allocated data structures. We have implemented this procedure in a prototype on top of existing SMT solvers and used it to automatically infer loop invariants of list-manipulating programs. This is joined work with Nishant Totla.

4 Working Groups

The goal of the seminar was to explore synergies between game theoretic synthesis algorithms and decision procedures. To make the potential for collaboration clear and concrete, we identified a number of research questions to be addressed during our seminar:

Tools for Systems Design Which tools and techniques developed for software/hardware verification can be used to engineer next generation industrial-strength synthesis systems?

Solvers for Concurrency What further theory extensions are needed to design efficient analysis and synthesis procedures for concurrent programs? For example, what reasoning capabilities are required for the synthesis of concurrent data structures? For synthesis of parameterized systems? For reasoning in the presence of weak memory models?

Games Modulo Theories How to solve games over rich domains, for example, over infinite alphabets subject to a background theory? Do decision procedures provide practical abstraction mechanisms for synthesis problems? What of more expressive properties, such as those studied in stochastic and quantitative games?

Synthesis and New Capabilities for Decision Procedures What new capabilities must decision procedures expose to help in synthesis? For example, can interpolants be used in generating certificates for synthesis problems? Which fixed-point logic is best solved with which search method and/or deductive techniques?

Games and Systems Composition How can we decompose large specifications into smaller specifications that can be synthesized automatically? What is the best way to use game-theoretic solutions for composable language design?

This Dagstuhl Seminar 12461 "Games and Decisions for Rigorous Systems Engineering" brought together renowned as well as young aspiring researchers from three groups.

- The first group was formed by researchers developing new paradigms of concurrent software, such as multi-core, distributed computing, and software testing.
- The second group comprised researchers working on reactive synthesis, such as decompositions of large specifications, parameterised synthesis, game theoretic models and partial-information game models.
- The third group consisted of researchers who design and combine decision procedures for various logical formalisms, such as propositional satisfiability (SAT), satisfiability modulo theory (SMT), quantified boolean formulas, and first-order logic.

5 Discussions

Our seminar initiated discussions between experts and young researchers of exceptional talent from reactive synthesis and automated deduction. Moreover, the presence of academia and industry enabled to discuss problems that are challenging both from the theoretical and practical point of view. These problems included the use of decision procedures and automated deduction in automata-based synthesis; generalizations of model checking algorithms within game-theoretic frameworks; integration of model checking with complementary techniques such as software testing; and extending SMT solvers and theorem provers with proof generation, unsatisfiable core extraction, and Craig interpolation.

6 Collaborations and Interaction

An outcome of the seminar and interaction is a collection of software model checking benchmarks in Horn format. These benchmarks have now been added to the Software Verification Competition – SVCOMP repository and are available from https://svn.sosy-lab.org/software/sv-benchmarks/trunk/clauses/. The benchmarks include checking safety assertions and are coming from the following sources:

- Boolean programs extracted from the SDV/SLAM research distribution;
- C programs from SVCOMP 2013, provided by Arie Gurfinkel;
- numeric programs from the ARMC model checker, provided by Andrey Rybalchenko;
- heap manipulating programs from the SLAyer tool, provided by Jael Kriener;
- driver programs, provided by Ken McMillan;
- Geometry design problems, provided by Zachary Kincaid;
- Liquid type checking problems from OCaml and Haskell, provided by Ranjit Jhala;
- Benchmarks from the Eldarica tool, provided by Hossein Hojjat, Philipp Rümmer, and Viktor Kuncak.

All together we collected around 10000 benchmarks published in the Horn format. The examples use domains ranging from Booleans, linear real arithmetic, linear integer arithmetic, and linear integer arithmetic combined with arrays. The benchmarks come from many different sources and problem domains, but share the characteristics that the queries amount to checking satisfiability of Horn clauses modulo theories. They are in the SMT-LIB interchange format, which is standardized on http://smtlib.org.

Participants

- Francesco Alberti
University of Lugano, CH
- Dirk Beyer
Universität Passau, DE
- Nikolaj Bjørner
Microsoft Res. – Redmond, US
- Tomas Brazdil
Masaryk University, CZ
- Krishnendu Chatterjee
IST Austria –
Klosterneuburg, AT
- Swarat Chaudhuri
Rice University, US
- Laurent Doyen
CNRS, ENS – Cachan, FR
- Uwe Egly
TU Wien, AT
- Azadeh Farzan
University of Toronto, CA
- Bernd Finkbeiner
Universität des Saarlandes, DE
- Alberto Griggio
Fondazione Bruno Kessler –
Trento, IT
- Aarti Gupta
NEC Laboratories America, Inc.
– Princeton, US
- Ashutosh Kumar Gupta
IST Austria –
Klosterneuburg, AT
- Arie Gurfinkel
CMU – Pittsburgh, US
- Jochen Hoenicke
Universität Freiburg, DE
- Radu Iosif
VERIMAG – Gières, FR
- Moa Johansson
Chalmers UT – Göteborg, SE
- Igor Konnov
TU Wien, AT
- Konstantin Korovin
University of Manchester, GB
- Laura Kovacs
TU Wien, AT
- Axel Legay
INRIA – Rennes, FR
- Rupak Majumdar
MPI for Software Systems –
Kaiserslautern, DE
- Anca Muscholl
Université Bordeaux, FR
- Albert Oliveras
TU of Catalonia – Barcelona, ES
- Joel Ouaknine
University of Oxford, GB
- David Parker
University of Birmingham, GB
- Brigitte Pientka
McGill Univ. – Montreal, CA
- Ruzica Piskac
MPI für Softwaresysteme –
Saarbrücken, DE
- Albert Rubio
UPC – Barcelona, ES
- Andrey Rybalchenko
TU München, DE
- Helmut Seidl
TU München, DE
- Martina Seidl
University of Linz, AT
- Natasha Sharygina
University of Lugano, CH
- Ana Sokolova
Universität Salzburg, AT
- Armando Solar-Lezama
MIT – Cambridge, US
- Eugenia Ternovska
Simon Fraser University –
Burnaby, CA
- Helmut Veith
TU Wien, AT
- Tomas Vojnar
Brno Univ. of Technology, CZ
- Andrei Voronkov
University of Manchester, GB
- Georg Weissenbacher
TU Wien, AT
- Josef Widder
TU Wien, AT
- Thomas Wies
New York University, US
- Florian Zuleger
TU Wien, AT

Report from Dagstuhl Seminar 12462

Symbolic Methods for Chemical Reaction Networks

Edited by

Francois Boulier[1]**, Anne J. Shiu**[2]**, Thomas Sturm**[3]**, and Andreas Weber**[4]

1 **Université de Lille I, FR,** `Francois.Boulier@univ-lille1.fr`
2 **University of Chicago, US,** `annejls@math.uchicago.edu`
3 **MPI für Informatik – Saarbrücken, DE,** `sturm@mpi-inf.mpg.de`
4 **Universität Bonn, DE,** `weber@cs.uni-bonn.de`

Abstract

During 11–16 November 2012, the Dagstuhl Seminar 12462 "Symbolic Methods for Chemical Reaction Networks" was held in Schloss Dagstuhl – Leibneiz Center for Informatics. The seminar brought together researchers in symbolic computation, chemical engineering, and systems biology. During the seminar, participants presented five-minute talks introducing their research interests, five participants gave longer talks, and all participants had the opportunity to take part in various discussion groups. Abstracts of presentations and summaries of the discussion groups are compiled in this report.

Seminar 11.–16. November, 2012 – www.dagstuhl.de/12462

1998 ACM Subject Classification G.1.5 Roots of Nonlinear Equations, G.1.7 Ordinary Differential Equations, I.1.4 Applications (Symbolic and Algebraic Manipulation), J.2 Physical Sciences and Engineering (Computer Applications), J.3 Life and Medical Sciences (Computer Applications)

Keywords and phrases chemical reaction network theory, differential algebra, Hopf bifurcation, mass-action kinetics, model inference, model reduction, molecular interaction network, multistationarity, parameter estimation, quantifier elimination, stoichiometric network analysis, symbolic computation

Digital Object Identifier 10.4230/DagRep.2.11.66

1 Executive Summary

Anne J. Shiu

Systems of differential equations and hybrid systems more generally are prevalent in chemical engineering and systems biology. The analysis of such systems focuses on resolving their dynamical properties, for instance, determining their equilibria and capacity for multistationarity or Hopf bifurcations. The additional tasks of parameter estimation, model reduction, and model inference are also relevant for these systems. These goals are difficult in general, especially due to the large size of these systems, especially those arising in systems biology. Non-numeric methods are essential in this context, because reaction parameters can vary over a wide range, parameter uncertainty is predominant in systems biology, and even the qualitative behavior of a system typically varies among different regions of the parameter space. Two major lines of research in this area are represented by chemical reaction network

Symbolic Methods for Chemical Reaction Networks, *Dagstuhl Reports*, Vol. 2, Issue 11, pp. 66–86
Editors: Francois Boulier, Anne J. Shiu, Thomas Sturm, and Andreas Weber
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

theory and stoichiometric network analysis. Our Dagstuhl seminar brought together researchers from both of these research areas, as well as researchers in symbolic computation and those on the application side (chemical engineering and systems biology). The aim of our seminar was twofold: to introduce practitioners to existing relevant theory and software from symbolic computation, and to allow participants to pose current computational challenges in this area, in order to spur development of symbolic computation methods to resolve these problems. To this end, collaborative working groups on various related topics were held throughout the week.

On Monday during the seminar, most of the participants gave short talks introducing their research interests. On Tuesday, long talks were given by Gheorghe Craciun and Francois Fages. A long talk on Wednesday was given by Stefan Schuster. Markus Eiswirth and Holger Fröhlich gave long talks on Thursday. Discussion groups met throughout the week.

2 Table of Contents

Executive Summary

Overview of Short Talks

Overview of Long Talks

Working Groups

3 Overview of Short Talks

This section contains abstracts of the short (three-slide, five-minute) talks.

3.1 Differential algebra: theory and applications

Francois Boulier (Université de Lille I, FR)

I talked about differential algebra: theory, applications to engineering, chemistry, biology, and the software I am developing.

3.2 Symbolic computational tools for understanding real polynomial equalities and inequalities

Christopher W. Brown (U.S. Naval Academy – Annapolis, US)

This talk gave a brief overview of the tools symbolic computation provides for understanding objects defined by real polynomial equalities and inequalities. Particular attention was paid to the idea of parametric versions of standard problems, and how solutions are often very different kinds of objects for parameterized problems.

3.3 Qualitative dynamics of biochemical reaction networks

Carsten Conradi (MPI – Magdeburg, DE)

I introduced some research problems currently under investigation in my research team "Qualitative Dynamics of Biochemical Reaction Networks" and briefly introduced two recent results concerning multistationarity and switching in mass action networks.

3.4 Symbolic methods for chemical reaction networks

Alicia Dickenstein (University of Buenos Aires, AR)

I surveyed my recent work on using methods from algebraic geometry for analyzing chemical reaction networks taken with mass-action kinetics. In particular, I discussed the topics of steady state invariants, multistationarity, and absolute concentration robustness.

3.5 A satisfiability-based approach to hybrid systems analysis

Andreas Eggers (Universität Oldenburg, DE)

This talk briefly summarizes the Satisfiability modulo Ordinary Differential Equations (SAT modulo ODE) approach to the analysis of hybrid systems. Starting for example from a hybrid automaton model, we build a predicative encoding of the transition system, which describes the discrete jumps and continuous flows. The conjunction of a predicate describing the system's possible initial states with unwindings of the transition predicate and a target predicate characterizing states whose reachability is of interest yields a SAT modulo ODE formula consisting of Boolean connectives, arithmetic atoms, and ODE constraints. Using our iSAT-ODE solver, we can find candidate solution boxes or prove the absence of solutions to formulae of this kind and thereby provide a tool for bounded model checking (BMC) of hybrid systems. We have also shown that unbounded properties like region stabilization can sometimes be encoded as BMC problems. The iSAT-ODE solver combines techniques from propositional SAT solving with Interval Constraint Propagation and enclosure methods for ODEs using the VNODE-LP solver and a bracketing system approach.

More details can be found in [1].

References

1 A. Eggers, N. Ramdani, N. S. Nedialkov, M. Fränzle. Improving the SAT modulo ODE approach to hybrid systems analysis by combining different enclosure methods. Software and Systems Modeling, Springer, 2012 to appear. DOI: 10.1007/s10270-012-0295-3.

3.6 Computing Hopf bifurcations in chemical reaction networks Using Reaction Coordinates

Hassan Errami (Universität Bonn, DE)

For low-dimensional reaction systems without additional constraints, Hopf bifurcation can be computed by reducing the question of its occurrence to quantifier elimination problems on real closed fields. However deciding its occurrence in high-dimensional systems has proven to be difficult in practice. I discussed a fully algorithmic technique to compute Hopf bifurcation fixed point for reaction systems with linear conservation laws using reaction coordinates instead of concentration coordinates, a technique that extends the range of networks which can be analyzed in practice, considerably.

3.7 The Contraintes project-team

Francois Fages (INRIA Le Chesnay, FR)

Contraintes[1] is a project-team of Inria[2], located in the Paris-Rocquencourt research center[3]. Created in March 2001, Contraintes investigates the theoretical foundations, design, implementation and applications of formal methods from computer science to mastering the complexity of complex systems in two domains: real-life combinatorial optimization problems and cell biology. Our four main scientific themes

- Rule-based Languages
- Constraint Solving Techniques
- Formal Methods for Systems Biology
- Integration of In Silico and In Vivo Approaches

are experimented through the development of the Biochemical Abstract Machine (BIOCHAM) modeling software and applied in collaboration with biologists to the building of computational models of cell signaling, gene expression and cell cycle control.

3.8 Multistationarity and variable elimination in reaction networks

Elisenda Feliu (University of Copenhagen, DK)

In this talk I briefly discussed two mathematical results about the properties of chemical reaction networks. The first result concerns the preclusion of multistationarity in reaction networks for which only qualitative information on the rate functions is known. The second result is a graph-based approach to decide what variables can be linearly eliminated from the steady-state equations of a mass-action system. Finally, applications to the study of different phosphorylation systems were discussed.

3.9 From biological networks to personalized medicine

Holger Fröhlich (Universität Bonn, DE)

This talk covered bioinformatics approaches to deal with biological networks. An overview about projects related to the workshop was presented.

[1] http://www.inria.fr/en/teams/contraintes
[2] http://www.inria.fr/en/
[3] http://www.inria.fr/en/centre/paris-rocquencourt

3.10 Symbolic analysis of finite difference approximations to differential equations describing chemical reaction systems

Vladimir Gerdt (JINR, LIT – Dubna, RU)

To simulate chemical processes one has to solve systems of ODEs (chemical reactions, control of nonlinear chemical processes), systems of DAEs (Rober's problem) or systems of PDEs (diffusion of chemicals on biological cells or membranes, pattern formations in biology, nonlinear chemical oscillators in excitable media). In most cases such systems are polynomially-nonlinear and can not be solved symbolically.

To solve systems of DEs the finite difference method is widely used. It is based upon the application of a local Taylor expansion to replace every differential equation by the difference one defined on the chosen computational grid. The obtained difference equations form a finite difference approximation (FDA) to the given DEs, and together with discrete approximation of initial or/and boundary conditions constitute a finite difference scheme.

An FDA to a system of DEs such that the latter inherits at the discrete level all the properties of the former is s(strongly)-consistent. For a uniform and orthogonal grid s-consistency admits symbolic verification [1]. For numerical solving ODEs or PDEs we suggest to use symbolic methods developed in [1] to select those FDA which are s-consistent.

References

1 V.P. Gerdt. Consistency analysis of finite difference approximations to PDE systems. LNCS 7175 (2012) 28–42. arXiv:math.AP/1107.4269

3.11 Functoriality of mass-action kinetics

Manoj Gopalkrishnan (TIFR Mumbai, IN)

Can mass-action kinetics be viewed as a functor from a category of reaction diagrams to a category of differential inclusions, and is this a profitable way to think of mass-action kinetics?

3.12 Finding steady points of mass-action kinetics

Dima Grigoriev (Université de Lille I, FR)

An algorithm is designed which finds steady points with non-zero coordinates of mass-action kinetics whose complexity is better than for existing ones when the number of linearly independent monomials in a system is small enough.

3.13 Symbolic methods for parameter estimation in chemical reaction networks

Daniel Kaschek (Universität Freiburg, DE)

Parameter estimation in chemical reaction networks often focuses on statistical methods which are implemented numerically. The use of symbolical methods simplifies numerical problems, makes the numerical methods more accurate and leads to parameter transformations turning the estimation problem into a better behaving problem. Examples include symbolic expressions for steady states and exploitation of Lie symmetries.

3.14 Is linear integer arithmetic useful?

Marek Kosta (MPI für Informatik – Saarbrücken, DE)

In automated reasoning there are different ways how to combine arithmetic with first-order reasoning. We pose the question: How could we use these combinations with linear integer arithmetic in the field of analysis of chemical reaction networks?

3.15 Computer algebra applied to chemical reaction systems

Francois Lemaire (Université de Lille I, FR)

I am specialized in Computer Algebra, especially the manipulation of algebraic and differential equations. In my short presentation, I explained how the techniques we develop could be applied in modeling in the context of chemical reaction systems. I expressed my motivation in finding new persons to collaborate with concerning the following directions: integrate our techniques in other people's software, help people analyze their chemical reaction systems, develop new algorithms for automating computations made by hand.

3.16 Symbolic methods for chemical reaction networks

Stefan Müller (RICAM – Linz, AT)

As a member of both a mathematics and a biotechnology institute, my research interests concerning "Symbolic Methods for Chemical Reaction Networks" are two-fold: one goal is to generalize CRNT (e.g. the deficiency zero theorem) for power-law kinetics which turns out to be relevant for reactions in intracellular environments, but also in purely chemical settings; the other aim is to simplify the modeling process (e.g. of a photobioreactor) or the bifurcation analysis (e.g. of a gene regulatory network) by automatizing the computation of (quasi) steady states.

3.17 Metabolic networks from enzyme's domains point of view

Sabine Peres (Université Paris Sud, FR)

In metabolic pathway analyses, the metabolic networks are described in term of biochemical reactions and metabolites. The integration of structural data is required for a comprehensive understanding of the metabolic networks. We represent the metabolic networks with the functional connectivity between the protein functional domains to make more relevant analyses. We used BioPsi, a formal multi-level description based on elementary actions, to assign functions on structural domains and the elementary flux modes theory to check if the already known pathways remain present and to identify new ones. This methodology is applied to the first part of the tricarboxylic acid cycle and the elementary flux modes considering protein domains reveals important aspects in its metabolic pathways.

3.18 Graph-based approaches for the analysis of biochemical regulation networks

Nicole Radde (Universität Stuttgart, DE)

Modeling the dynamics of intracellular regulation networks by systems of ordinary differential equations has become a standard method in systems biology, and it has been shown that the behavior of these networks is often tightly connected to the network topology. We have recently introduced the circuit-breaking algorithm (CBA), a method that uses the network topology to construct a one-dimensional circuit-characteristic of the system. It was shown that this characteristic can be used for an efficient calculation of the system's fixed points. This work was further extended by showing several connections between the circuit-characteristic and the stability of fixed points.

My presentation at the Dagstuhl seminar will focus on graph-based approaches for biological regulation networks, with a focus on the CBA. All statements are illustrated on biological network models.

3.19 Generalized mass action systems: Complex balancing equilibria and sign vectors of the stoichiometric and kinetic-order subspaces

Georg Regensburger (RICAM – Linz, AT)

Mass action systems capture chemical reaction networks in homogeneous and dilute solutions. We suggest a notion of generalized mass action systems that admits arbitrary nonnegative power-law rate functions and serves as a more realistic model for reaction networks in intracellular environments. In addition to the chemical complexes and the related stoichiometric subspace, we introduce corresponding kinetic complexes, which represent the nonnegative exponents in the rate functions and determine the kinetic-order subspace. We show that several results of Chemical Reaction Network Theory carry over to the case of generalized mass action kinetics. Our main result essentially states that, if the sign vectors of the stoichiometric and the kinetic-order subspace coincide, there exists a unique positive complex balancing equilibrium in every stoichiometric compatibility class. However, in contrast to classical mass action systems, multiple complex balancing equilibria in one stoichiometric compatibility class are possible in general.

3.20 PoCaB: A software infrastructure to explore algebraic methods for bio-chemical reaction networks

Satya Swarup Samal (B-it – Bonn, DE)

Given a bio-chemical reaction network, I discussed the different algebraic entities e.g. stoichiometric matrix, polynomial system, deficiency and flux cones which are prerequisite for the application of various algebraic methods to qualitatively analyze them. These entities are computed on the examples obtained from two publicly available bio-databases called Biomodels and KEGG and stored in a publicly available database called PoCaB (http://pocab.cg.cs.uni-bonn.de/).

3.21 SMT solving for hybrid systems

Karsten Scheibler (Universität Freiburg, DE)

Over the past decades embedded systems have become more and more complex. Furthermore, they are now often a combination of digital components and analog parts – making them to embedded hybrid systems. Especially in safety critical environments, a formal correctness proof of such systems is highly desirable. The SMT solver iSAT is suitable to verify safety properties of systems consisting of both, linear and non-linear behaviour. Because of its ability to handle arbitrary boolean combinations of linear and non-linear constraint formulas, iSAT is also a natural choice for solving such formulas arising in the area of chemical reaction networks.

3.22 Stability and bifurcation analysis of major reaction subnetworks in three-way catalytic monoliths

Igor Schreiber (Institute of Chemical Technology – Prague, CZ)

The three-way catalytic monolithic converter (TWC) used for automobile emission control is the most common reactor. The processes in the TWC include mass transport coupled with simultaneous heterogeneous catalytic oxidation of CO and hydrocarbons, and reduction of NOx in a monolith – a multichannel flow-through reactor. In the first step, we focus on the analysis of complex dynamical modes stemming from the complexity of the chemical and adsorption processes involved. Therefore we take the simplest approximation of the system – an isothermal continuous stirred tank reactor (CSTR). The model exhibits various types of dynamics characteristic of nonlinear systems including multiple steady states, oscillations and chaos. These dynamical features are found via bifurcation diagrams constructed by numerical continuation techniques. Such phenomena are caused by interactions among various reaction and adsorption steps, not by the thermokinetic effect. To account for these dynamical modes, we analyze the reaction mechanism taken from the literature with the use of the stoichiometric network analysis – a tool for decomposition of the reaction network and determination of unstable subnetworks. Such reaction subnetworks may possess positive and negative feedback loops, which may imply periodic oscillations. This analysis is crucial for interpretation of the chemical nature of oscillations observed within certain region of parameters in the oxygen inflow vs. temperature bifurcation diagrams.

3.23 Symbolic computations for control and system theory of biochemical reaction systems

Jan H. van Schuppen (VSCR – Amsterdam, NL)

Biochemical reaction systems are mostly polynomial or rational systems with inputs and outputs as considered in system theory. System identification and system reduction algorithms often require the knowledge of whether a system satisfies particular system theoretic properties and those properties are best determined by symbolic computations. Controllability, observability, and structural identifiability are system properties for which symbolic computations are needed. The lecture contains examples of biochemical reaction systems and a summary of system theory of rational systems. The lecturer hopes to raise the interest of researchers in symbolic computations for the problems of differential algebra of rational systems and of biochemical reaction systems.

3.24 General systems of differential equations

Werner M. Seiler (Universität Kassel, DE)

I briefly review the notion of a "general system" of differential equations. The first part discusses the dichotomy between normal systems and those that are under- or overdetermined and emphasizes the need for a completion to involution. The second part is concerned with implicit systems and introduces several types of singular behavior. In the context of chemical reaction networks, the first topic is related for example to problems arising in a quasi steady state approximation or to the treatment of conservation laws, whereas the second topic provides an alternative point of view of static bifurcations and also leads to phenomena like singularity induced bifurcations.

3.25 Multistationarity and persistence in chemical reaction networks

Anne J. Shiu (University of Chicago, US)

This talk presented two theorems, one concerning multistationarity and the other concerning persistence. The first result (due to joint work with Badal Joshi) states that multistationarity of certain networks can be inferred from the existence of multistationary "embedded" networks, that is, smaller networks obtained from the original network by removing chemical species or reactions. We then posed the challenge of enumerating the multistationary networks that are minimal with respect to this "embedded" network relation. The second result (due to joint work with Manoj Gopalkrishnan and Ezra Miller) states that "strongly endotactic" networks (informally, networks with reactions that point inward) are persistent, that is, trajectories with positive initial condition remain bounded away from zero in all coordinates. This theorem makes progress toward a conjecture of Craciun, Nazarov, and Pantea that asserts that so-called "endotactic" networks are persistent.

3.26 Real quantifier elimination

Thomas Sturm (MPI für Informatik – Saarbrücken, DE)

We explain the notion of real quantifier elimination and discuss by means of an example its sussessful application to the decision of the existence of Hopf bifurcations in a gene regulatory network. We point at the more general and more difficult problem to find necessary and sufficient conditions in the parameters for the existence of such bifurcation and suggest to discuss its relevance.

3.27 Chemical reaction networks: equilibria, Hopf bifurcations, and algorithms

Andreas Weber (Universität Bonn, DE)

For large-scale systems arising from chemical reaction networks, already the parametric computation of equilibria and singularities in the positive orthant is a major challenge—but also of great importance for applications by understanding properties of the reaction networks. We have implemented a newly developed algorithm by Dima Grigoriev, by which we could solve equilibria for most examples from the BIOMOD database and also for the examples showing "absolute concentration robustness" (Shinar and Feinberg, 2010).

As the existence of Hopf bifurcations points yield oscillations we build on previous work with M. El Kahoui and T. Sturm that yield reductions to quantifier elimination over the reals (and computations using REDLOG). In recent joint work with H. Errami, W. Seiler, M. Eiswirth we could compute larger examples fully algorithmically using reaction coordinates.

On the systems side we have built PoCaB – a platform to explore bio-chemical reaction networks by algebraic methods. It not only provides the software framework for automated computations involving different systems, but also contains a database for the algebraic entities computed from the models of chemical reaction networks.

4 Overview of Long Talks

This section contains abstracts of the five long talks.

4.1 Chemical reaction network theory: classical results and recent advances

Gheorghe Craciun (University of Wisconsin Madison, US)

Parameter-free analysis of biochemical reaction network models can often use methods from chemical reaction network theory (CRNT), which was created in the 1970s by Horn, Jackson and Feinberg. We describe some classical results of CRNT and discuss recent extensions of this theory for understanding multistability, persistence, oscillations, and chaotic dynamics in biochemical reaction network models, and in general mathematical models of biological interaction networks. These models lead to polynomial dynamical systems for which functional properties can be analyzed using large-scale symbolic computation.

4.2 Stiochiometric network analysis

Ralf Markus Eiswirth (FHI – MPG Berlin, DE)

By making maximum use of stoichiometric restrictions, Stoichiometric Network Analysis (SNA) allows to solve for the complete set of stationary states of chemical reaction systems and write it down in closed form as a linear combination of extreme subnetworks. Consequently, it is possible to decide whether an instability can occur in a given network anywhere in parameter space. The procedure will be outlined and a classification of chemical instabilities will be given.

4.3 Inferring reaction models from ODEs and model reductions as subgraph epimorphisms

Francois Fages (INRIA Le Chesnay, FR)

Many models in Systems Biology are described as Ordinary Differential Equations (ODEs), which allow for numerical integration, bifurcation analyses, parameter sensitivity analyses, etc. However, before fixing the kinetics and parameter values and going to simulations, various analyses can be performed based only on the structure of the model. This approach has rapidly developed in Systems Biology in the last decade, with for instance, the analyses of structural invariants in Petri net representation [3], model reductions by subgraph epimorphims [1], qualitative attractors in logical dynamics or temporal logic properties by analogy to circuit and program verification. These complementary analysis tools do not rely on kinetic information, but on the structure of the model with reactions.

The Systems Biology Markup Language (SBML) of [2] is now a standard for sharing and publishing reaction models. However, since SBML does not enforce any coherence between the structure and the kinetics of a reaction, an ODE model can be transcribed in SBML without reflecting the real structure of the reactions, hereby invalidating many structural analyses.

In this talk we propose a general compatibility condition between the kinetic expression and the structure of a reaction, describe a symbolic computation algorithm for inferring a reaction model from an ODE system, and report on its use for automatically curating the writing in SBML of the model repository biomodels.net. We illustrate the benefit of this curation by computing faithful hierarchies of models in biomodels.net, related by model reduction relationships defined solely on the structure of the reaction graphs as subgraph epimorphisms.

References

1 S. Gay, S. Soliman, F. Fages. A graphical method for reducing and relating models in systems biology. Bioinformatics 26:18 (2010) i575–i581. Special issue ECCB'10.

2 M. Hucka et al. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. Bioinformatics, 19:4 (2003) 524–531. URL http://sbml.org/.

3 V.N. Reddy, M.L. Mavrovouniotis, M.N. Liebman. Petri net representations in metabolic pathways. In L. Hunter, D.B. Searls, and J.W. Shavlik, editors, Proceedings of the 1st International Conf. Intell. Syst. Mol. Biol., 1993.

4.4 Quantitative and qualitative dynamic modeling of molecular interaction networks

Holger Fröhlich (Universität Bonn, DE)

Molecular interaction networks, such as metabolic networks, signaling pathways and gene regulatory networks, are not static, but have to be understood as dynamical systems. Various computational methods have been established in the literature to model these systems in order to obtain experimentally verifiable hypotheses. Besides quantitative approaches, which are typically formulated in terms of ordinary differential equation systems (ODEs), Boolean Networks have gained a lot of attention for this purpose. This lecture will highlight differences as well as connections between these two modeling schemes and discusses advantages and disadvantages of both from a practical point of view. In this context I will also shed light on the problem of estimating kinetic rate parameters from experimental data via different statistical techniques.

4.5 Unraveling the structure of metabolism by elementary-modes analysis

Stefan Schuster (Friedrich-Schiller-Universität – Jena, DE)

Cellular metabolism has a complex structure due to the large number of reactions involved and the fact that many of these are bi- or trimolecular. To unravel this structure, elementary-mode analysis has become a well-established theoretical tool. It allows one to decompose complex metabolic networks into the smallest functional entities, which can be interpreted as biochemical pathways or, as a special case, substrate cycles. Moreover, it allows the maximization of molar yields, which has important applications in biotechnology. Here, we outline the theoretical basis and the central concepts and algorithms in elementary-mode analysis. Moreover, several illustrative examples of application of that analysis are presented. Current trends and future prospects are discussed.

References

1 L.F. de Figueiredo, T.I. Gossmann, M. Ziegler, S. Schuster. Pathway analysis of NAD+ metabolism. Biochem. J. 439 (2011) 341–348.

2 J. Gebauer, S. Schuster, L. F. de Figueiredo, C. Kaleta. Detecting and investigating substrate cycles in a genome-scale human metabolic network. FEBS J. 279 (2012) 3192–3202.

3 E. Ruppin, J.A. Papin, L.F. de Figueiredo, S. Schuster. Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. Curr. Opin. Biotechnol. 21 (2010) 502–510.

4 S. Schuster, D.A. Fell, T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nature Biotechnol. 18 (2000) 326–332.

5 Working Groups

This section contains summaries of each of the discussion groups.

5.1 Bistability

Carsten Conradi

Recent results concerning reaction networks with generalized mass action kinetics obtained by Mueller and Regensburger were discussed. These results extend the notion of complex balanced equilibrium from networks with 'classical' mass action kinetics to networks allowing a more general form of mass action kinetics. Under certain assumptions on the network structure and the (generalized) kinetics one obtains complex balanced equilibria. In this generalized setting uniqueness of complex balanced equilibria is no longer guaranteed and a sufficient condition for multiple equilibria in terms of sign patterns was discussed.

This discussion group met on Wednesday–Friday, and participants included Carsten Conradi, Markus Eiswirth, Elisenda Feliu, Stefan Müller, Georg Regensburger, Jan van Schuppen, Igor Schreiber, Anne Shiu.

5.2 Preprocessing methods for systems of real constraints

Thomas Sturm

The essential complexity parameter for existential decision problems over the reals is the number of variables n in contrast to the number of constraints k. Dima Grigoriev has published theoretical complexity results for the special case of quadratic constraints, which comprise a satisfiability-preserving transformation for decreasing the number of variables. Essentially, after the transformation there are $2k$ variables instead of n, where k is the number of constraints in the input problem. It turns out that the existing real decision problems arising from chemical reaction systems contain few constraints in comparison to the number of variables; i.e. $k \ll n$. This makes an implementation and its application to chemical reaction systems appear quite appealing. On the other hand most of these examples have high degrees so that it is important to understand to what extent the existing approach could be generalized. Furthermore, although the existing result for the quadratic case is of algorithmic nature it is not straightforward to create corresponding software. Throughout the technical part of our discussions we analyzed the procedure both from a theoretical and from a software engineering point of view.

The discussion group met on both Thursday and Friday, and participants were Christopher Brown, Andreas Eggers, Dima Grigoriev, Marek Kosta, Karsten Scheibler, Thomas Sturm, and Andreas Weber.

5.3 Specialized decision procedures for the existential fragment of the reals

Thomas Sturm

Algebraic decision procedures, which recently have been successfully applied in the analysis of biological and chemical reaction networks, are typically based on quantifier elimination procedures, which in fact solve a more general problem by possibly deciding parametric problems. After translation to real algebra, relevant questions arising from chemical reaction systems often have a special form: They do not involve parameters, and they are purely existential. Furthermore, there are often natural bounds on the possible values of variables, e.g., many variables can be constrained to taking only positive values.

We discussed perspectives for more efficient specialized procedures for these situations. We identified as promising directions recent ideas by de Moura et al. for decision procedures resembling SMT approaches by constructing model assumptions and the use of interval arithmetic in contrast to computation with exact algebraic numbers.

This discussion group met on both Tuesday and Wednesday, and the participants were Christopher Brown, Andreas Eggers, Marek Kosta, Karsten Scheibler, and Thomas Sturm.

5.4 Conservation laws

Francois Lemaire

We discussed about the problem of defining and finding “good” conservation laws of a system of chemical reactions. A basis of conservation laws is easily computed as a kernel of a matrix. However, certain conservation laws are better than others. In particular, conservation laws with positive coefficient and the most zeros are particularly interesting. The outcome of the discussion was that applying techniques for computing elementary flux modes might help. Still, this approach might not be the best one since some combinatorial explosion might happen.

This discussion group met on Tuesday.

5.5 Model reduction by systematic exploitation of scaling symmetries

Daniel Kaschek

Francois Lemaire presented a software package that computes explicit coordinate and parameter transformations originating from scaling invariances of a system of ordinary differential equations (ODEs). Holger Fröhlich, Daniel Kaschek, and Nicole Radde provided typical examples of ODEs from systems biology projects. Possible applications to this class of systems have been discussed. In particular, the concrete structure of a given observation function has been identified to play a crucial role for the utility of the method. A number of new questions have arisen from the discussion, e.g. connections between polynomial conserved quantities and translational symmetries or polynomial symmetries and related coordinate and parameter transformations.

This discussion group met on Tuesday, and the participants were Holger Fröhlich, Daniel Kaschek, Francois Lemaire, and Nicole Radde.

5.6 Parameter estimation

Francois Boulier

Holger Fröhlich recalled the classical numerical approaches for this problem (frequentist and Bayesian approaches). Jan van Schuppen exposed the procedure which is traditional, in control theory, for tackling it. Francois Boulier presented differential algebra techniques related to this question together with a new algorithm which might help transforming differential equations into integral ones and thereby be helpful in the presence of noise. Nicole Radde presented a real problem she is concerned with and Daniel Kaschek started to try to solve it using his methods. Jan van Schuppen exposed computer algebra problems he would like to solve, related to observability, input design and system reduction. Gheorghe Craciun talked about theoretical difficulties which arise when trying to infer chemical reaction systems from the ODE models. However it was poined out that these difficulties might be sometimes turned into advantages. There were various questions by Andreas Weber and Werner Seiler.

This discussion group met on Tuesday and Wednesday.

Participants

- Francois Boulier
Université de Lille I, FR
- Christopher W. Brown
U.S. Naval Academy – Annapolis, US
- Carsten Conradi
MPI – Magdeburg, DE
- Gheorghe Craciun
University of Wisconsin – Madison, US
- Alicia Dickenstein
University of Buenos Aires, AR
- Andreas Eggers
Universität Oldenburg, DE
- Ralf Markus Eiswirth
FHI – MPG Berlin, DE
- Hassan Errami
Universität Bonn, DE
- François Fages
INRIA Le Chesnay, FR
- Elisenda Feliu
University of Copenhagen, DK
- Holger Fröhlich
Universität Bonn, DE
- Vladimir Gerdt
JINR, LIT – Dubna, RU
- Manoj Gopalkrishnan
TIFR Mumbai, IN
- Dima Grigoriev
Université de Lille I, FR
- Daniel Kaschek
Universität Freiburg, DE
- Marek Kosta
MPI für Informatik – Saarbrücken, DE
- Francois Lemaire
Université de Lille I, FR
- Stefan Müller
RICAM – Linz, AT
- Sabine Peres
Université Paris Sud, FR
- Adrien Poteaux
Université de Lille I, FR
- Nicole Radde
Universität Stuttgart, DE
- Georg Regensburger
RICAM – Linz, AT
- Satya Swarup Samal
B-it – Bonn, DE
- Karsten Scheibler
Universität Freiburg, DE
- Igor Schreiber
Institute of Chemical Technology – Prague, CZ
- Stefan Schuster
Friedrich-Schiller-Universität – Jena, DE
- Werner M. Seiler
Universität Kassel, DE
- Anne J. Shiu
University of Chicago, US
- Thomas Sturm
MPI für Informatik – Saarbrücken, DE
- Jan H. van Schuppen
VSCR – Amsterdam, NL
- Andreas Weber
Universität Bonn, DE

Report from Dagstuhl Seminar 12471

SAT Interactions

Edited by

Nadia Creignou[1], Nicola Galesi[2], Oliver Kullmann[3], and Heribert Vollmer[4]

1 **Université de Marseille, FR,** `creignou@lif.univ-mrs.fr`
2 **University of Rome "La Sapienza", IT,** `galesi@di.uniroma1.it`
3 **Swansea University, GB,** `o.kullmann@swansea.ac.uk`
4 **Leibniz Universität Hannover, DE,** `vollmer@thi.uni-hannover.de`

Abstract

This report documents the programme and outcomes of Dagstuhl Seminar 12471 "SAT Interactions". The seminar brought together researchers from different areas from theoretical computer science as well as the area of SAT solvers. A key objective of the seminar has been to initiate or consolidate discussions among the different groups for a fresh attack on one of the most important problems in theoretical computer science and mathematics.

Seminar 18.–23. November, 2012 – www.dagstuhl.de/12471

1998 ACM Subject Classification D.2.8 Complexity measures, F.2 Analysis of Algorithms and Problem Complexity, F.1.3 Complexity Measures and Classes, F.2.2 Nonnumerical Algorithms and Problems, F.4 Mathematical Logic and Formal Languages, G.2.1 Combinatorics

Keywords and phrases satisfiability problem, computational complexity, P-NP question, proof complexity, combinatorics, SAT-solvers, quantified Boolean formulas

Digital Object Identifier 10.4230/DagRep.2.11.87

Edited in cooperation with Johannes Ebbing

1 Executive Summary

Nadia Creignou
Heribert Vollmer

Brief Introduction to the Topic

Propositional satisfiability (or Boolean satisfiability) is the problem of determining whether the variables of a Boolean formula can be assigned truth values in such a way as to make the formula true. The satisfiability problem, SAT for short, stands at the crossroads of logic, graph theory, computer science, computer engineering and computational physics.

In particular SAT is of central importance in various areas of computer science including algorithmics, verification, planning and hardware design. It can express a wide range of combinatorial problems as well as many real-world ones. Due to its potential practical implications an intensive search has been done on how one could solve SAT problems in an automated fashion. The last decade has seen the development of practically-efficient algorithms for SAT, which can solve huge problems instances.

At the same time SAT is very significant from a theoretical point of view. Since the Cook-Levin's theorem, which has identified SAT as the first NP-complete problem, it has become a reference for an enormous variety of complexity statements. The most prominent one is the question "is P equal to NP?" Proving that SAT is not in P would answer this question negatively. Indeed, as stated by Richard Lipton on his blog *Gödel's Lost Letter and P = NP* (http://rjlipton.wordpress.com) such a proof matters since it would tell us why some computational problems are intrinsically more difficult than others, it would suggest new methods that would yield new insights on the fundamental nature of computation and it would help with goals of security for cryptographers.

During the past two decades, an impressive array of diverse techniques from mathematical fields, such as propositional logic, model theory, Boolean function theory, combinatorics, probability, and statistical physics has contributed to a better understanding of the SAT problem. Although significant progress has been made on several fronts, most of the central questions remain unsolved so far. One of the main aims of the Dagstuhl Seminar was to bring together researchers from different areas of activity in SAT (with an emphasize on mathematical aspects), so that they can communicate state-of-the-art advances and embark on a systematic interaction that will enhance the synergy between the different areas.

Organization of the Seminar and Activities

The workshop brought together 44 researchers from different areas of computer science and mathematics such as logic, complexity theory, algorithms, and proof complexity with complementary expertise. 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-five talks of various lengths took place over the five days of the workshop. Introductory and tutorial talks of 60 minutes were scheduled prior to workshop. Most of the remaining slots were filled, mostly with shorter talks, as the workshop commenced. The organizers considered it important to leave ample free time for discussion.

The tutorial talks were scheduled during the beginning of the week in order to establish a common background for the different communities that came together for the workshop. The presenters and topics were:

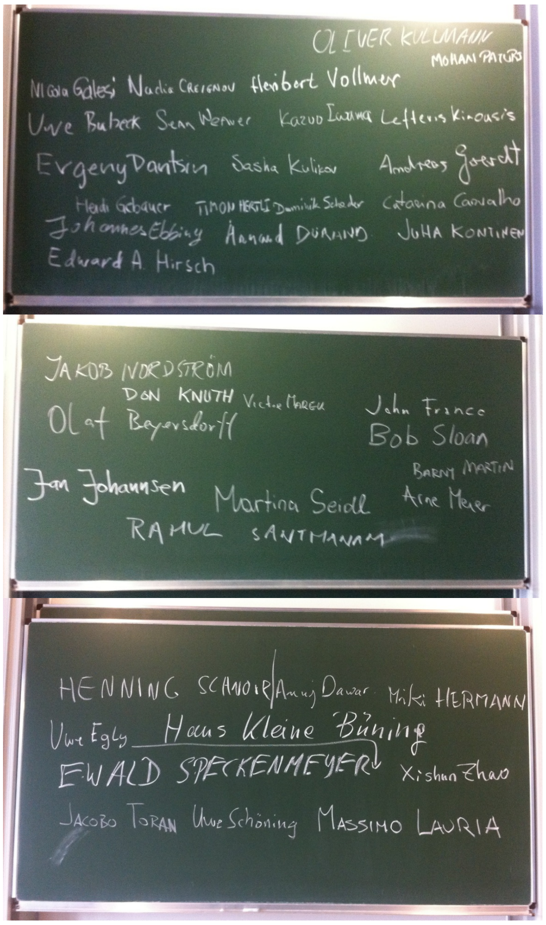
- Olaf Beyersdorff, Proof complexity
- Arne Meier, Complexity classifications for different satisfiability problems
- Victor Marek, Erdős' dream; SAT and combinatorics
- Uwe Bubeck, Quantified Boolean formulas: complexity and expressiveness
- Oliver Kullmann, The combinatorics of minimal unsatisfiability
- Martina Seidl, QBF solvers

Most of the tutorials were given by young researchers, reflecting the fact that the SAT community is dynamic and fast evolving.

A highlight of the seminar was the talk by Donald E. Knuth, delivered Wednesday morning, on "Satisfiability and the Art of Computer Programming". Knuth reported about his experiences while working on a chapter on satisfiability for the upcoming volume of his world-renowned series.

There were additionally 19 shorter talks. These talks covered a wide range of topics related to satisfiability. The different approaches discussed above in the seminar description were all very well represented by the different talks given during the five days of the seminar.

1. Combinatorics
   - Xishun Zhao, Finiteness conjecture on hitting minimal unsatisfiable formulas
   - Uwe Schöning, Probability distributions for local search and make versus break
   - Heidi Gebauer, Applications of (k, d)-trees


OLIVER KULLMANN
MOHAN PATURI
Nicola Galesi Nadia CREIGNOU Heribert Vollmer
Uwe Bubeck Sean Weaver Kazuo Iwama Lefteris Kirousis
Evgeny Dantsin Sasha Kulikov Andreas Goerdt
Heidi Gebauer TIMON HERTLI Dominik Scheder Catarina Carvalho
Johannes Ebbing Arnaud DURAND JUHA KONTINEN
Edward A. Hirsch
JAKOB NORDSTRÖM
DON KNUTH Victor MAREK
John Franco
Olaf Beyersdorff
Bob Sloan
BARNY MARTIN
Jan Johannsen
Martina Seidl
Arne Meier
RAHUL SANTHANAM
HENNING SCHNOOR Anuj Dawar Miki HERMANN
Uwe Egly Hans Kleine Büning
EWALD SPECKENMEYER Xishun Zhao
JACOBO TORAN Uwe Schöning MASSIMO LAURIA

2. Complexity
   - Juha Kontinen, Dependence logic and complexity
   - Julian-Steffen Müller, A fragment of dependence logic characterizing PTIME
   - Alexander Kulikov, New lower and upper bounds for Boolean circuit complexity
   - Johannes Ebbing, Model checking for modal intuitionistic dependence logic
3. Proof complexity
   - Uwe Egly, Proof complexity for QBF
   - Jan Johannsen, Separating clause learning proof systems from (regular) resolution
   - Jakob Nordström, Relating proof complexity measures and practical hardness of SAT
   - Massimo Lauria, Open problems in proof complexity
4. Algorithms
   - Stefan Szeider, Fixed-parameter tractability and SAT
   - Mohan Paturi, Algorithmic expressivity and hardness of satisfiability
   - Dominik Scheder, Exponential lower bounds for the PPSZ k-SAT algorithm

5. Logic
   - Arnaud Durand, A criterion for tractability of counting solutions to uniform CSP
   - Hans Kleine Büning, On some configuration problems based on representations in propositional logic
6. Solvers
   - John Franco, Adding unsafe constraints to improve satisfiability performance
   - Sean Weaver, Satisfiability enhancements enabled by state machines

This classification is necessarily rough, as several talks crossed the boundaries between these areas, in keeping with the theme of the workshop. The broad scope of the talks extended even to areas not anticipated by the organizers, such as dependence logic. The workshop thus achieved its aim of bringing together researchers from various related communities to share state-of-the-art research.

Concluding Remarks and Future Plans

The organizers regard the workshop as a great success. Bringing together researchers from different areas of theoretical computer science fostered valuable interactions and led to fruitful discussions. Feedback from the participants was very positive as well. Many attendants expressed their wish for a continuation and stated that this seminar was among the most fruitful Dagstuhl seminars they attended.

Finally, the organizers wish to express their gratitude toward the Scientific Directorate of the Center for its support of this workshop, and hope to establish a series of workshops on *SAT Interactions* in the future.

2 Table of Contents

3 Overview of Talks

3.1 Proof Complexity

Olaf Beyersdorff (University of Leeds, GB)

This talk surveys important results from the area of propositional proof complexity. In the talk I will highlight motivations and applications of proof complexity and important techniques which have been developed to show lower bounds. In particular, I will explain a game-theoretic technique which characterises tree-like Resolution size and illustrate this technique by proving the optimal lower bound for the pigeonhole principle in tree-like Resolution.

3.2 Quantified Boolean Formulas: Complexity and Expressiveness

Uwe Bubeck (Universität Paderborn, DE)

Joint work of Bubeck, Uwe; Kleine Büning, Hans
Main reference U. Bubeck, “Model-Based Transformations for Quantified Boolean Formulas,” in: Dissertations in Artificial Intelligence (DISKI), Vol. 329, Wolfgang Bibel (Ed.), IOS Press, ISBN 978-1-60750-545-7, 2010.
URL http://www.ub-net.de/bubeck-qbf-transformations-2010.pdf

We consider quantified Boolean formulas with free variables (QBF*) as an elegant way to represent Boolean functions. In the talk, we give an overview of fundamental concepts and complexity results, and we present suitable encoding techniques to compress propositional formulas by applying quantification.

We also relate QBF* to other representations of Boolean functions. In particular, we discuss the close relationship between existential quantification and Boolean circuits with unbounded fan-out [1, 2], as well as transformations between quantified Boolean formulas and nested Boolean functions (NBF) in both directions [3].

References

1 S. Aanderaa and E. Börger. *The Horn Complexity of Boolean Functions and Cook's Problem.* Proc. 5th Scandinavian Logic Symposium 1979, Aalborg University Press, 1979

2 H. Kleine Büning, X. Zhao, and U. Bubeck. *Resolution and Expressiveness of Subclasses of Quantified Boolean Formulas and Circuits.* Proc. 12th Intl. Conf. on Theory and Applications of Satisfiability Testing (SAT 2009), Springer, 2009

3 U. Bubeck and H. Kleine Büning. *Encoding Nested Boolean Functions as Quantified Boolean Formulas.* Journal on Satisfiability, Boolean Modeling and Computation (JSAT) 8(1): 101-116, 2012

3.3 Structural Tractability of Counting of Solutions to Conjunctive Queries

Arnaud Durand (University Paris-Diderot, FR)

Joint work of Durand, Arnaud; Mengel, Stefan

This talk survey some recent characterization results obtained on the counting complexity of subclasses of conjunctive queries (i.e. constraint satisfaction problem with projection). We prove that for counting of acyclic conjunctive queries (and many more fragments) it is possible to chart the tractability frontier. One of the main ingredients of this characterization is a new parameter associated to formulas that measure how free variables are spread into formulas.

3.4 Model Checking for Modal Intuitionistic Dependence Logic

Johannes Ebbing (Leibniz Universität Hannover, DE)

Joint work of Ebbing, Johannes; Lohmann, Peter; Yang, Fan

Modal intuitionistic dependence logic (MID) incorporates the notion of "dependence" between propositions into the usual modal logic and has connectives which correspond to intuitionistic connectives in a certain sense. It is the modal version of a variant of first-order dependence logic introduced by Väänänen in [1] considered by Abramsky and Väänänen [2] basing on Hodges' team semantics (1997). In this talk we give an overview on the computational complexity of the model checking problem for MID and its fragments built by restricting the operators allowed in the logics. In particular, we see that the model checking problem for MID is in general PSPACE-complete and that for propositional intuitionistic logic is coNP-complete.

References

1 J. Väänänen, *Dependence logic: A new approach to independence friendly logic*, London Mathematical Society student texts, no. 70, Cambridge University Press, 2007.

2 S. Abramsky and J. Väänänen, *From IF to BI*, Synthese 167 (2009), no. 2, 207–230.

3.5 Adding Unsafe Constraints to Improve the Performance of SAT Algorithms

John Franco (University of Cincinnati, US)

Joint work of Franco, John; Kouril, Michal
Main reference M. Kouril, J. Franco, "Resolution Tunnels for Improver SAT Solver Performance," in Proc. of the 8th Int'l Conf. on Theory and Applications of Satisfiability Testing (SAT'05), LNCS, Vol. 3569, pp, 143–158, 2005.
URL http://dx.doi.org/10.1007/11499107_11

For many families of SAT formulas the difficulty in solving an instance escalates exponentially with increasing instance size. A possible reason for this is that inferred constraints that reduce search space significantly are learned too late in the search to be effective. One attempt to control this is to add safe, uninferred constraints that are obtained from an analysis of the problem or the structure of the formula: for example symmetry breaking constraints. This approach proves effective in some but not all cases. We propose an alternative approach which is to add unsafe, uninferred constraints early on to reduce search space breadth at shallow depth and then retract those constraints when the search breadth is still small and will not get much bigger as search continues. By "unsafe constraint" we mean a constraint that may eliminate one or more satisfying assignments – hence there is a risk that all assignments of satisfiable instance may be eliminated.

We show, for example that in the case of formulas for solving van der Waerden number W(2,6), adding unsafe constraints produces a bound that turns out to be W(2,6). Knowledge of this bound and the conjecture that it was W(2,6) was eventually used by Kouril to custom design a solver that could prove definitively the value of W(2,6). Notable is the fact that the unsafe constraints are obtained from an analysis of solutions to smaller instances of the van der Waerden family and not from an analysis of the structure of the formulas or problem properties.

3.6 On Configuration Problems Based on Representations in Propositional Logic

Hans Kleine Büning (Universität Paderborn, DE)

Joint work of Bubeck, Uwe; Kleine Büning, Hans; Yan, Yuhan; Zhao, Xishun

We consider configuration problems, where the components are represented by propositional formulas. Configuration is the process of composing a system from a predefined set of components, while observing a set of given constraints and customer demands.

We focus on the computational complexity of the configuration problem for various sub-classes of formulas and architectures of the desired system. For example, the desired architecture can be a set of components or a circuit whose nodes are components computing Boolean functions

Moreover, we investigate the so-called specification problem in which we want to learn an unknown component given a partial solution of the configuration problem.

3.7 New Lower and Upper Bounds for Boolean Circuit Complexity

Alexander S. Kulikov (Steklov Inst. – St. Petersburg, RU)

Joint work of Demenkov, Evgeny; Kojevnikov, Arist; Kulikov, Alexander S.; Melanich, Olga; Mihajlin, Ivan; Yaroslavtsev, Grigory

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.

3.8 The Combinatorics of Minimal Unsatisfiability

Oliver Kullmann (Swansea University, GB)

URL http://cs.swan.ac.uk/ csoliver/papers.html#MU2012Dagstuhl

A talk giving an overview on the project of classifying minimally unsatisfiable clause-sets. The basic intuitions behind this project are:

- unsatisfiability of clause-sets can be "explained" by the included minimally unsatisfiable clause-sets
- minimally unsatisfiability can be reduced to basic, intuitive patterns, when using the deficiency (the difference between the number of clauses and the number of variables) as complexity parameter.

The fundamentals are discussed, and then new results, not included in the chapter in the Handbook of Satisfiability, are outlined.

References

1 Armin Biere, Marijn J.H. Heule, Hans van Maaren, and Toby Walsh, editors. *Handbook of Satisfiability*, volume 185 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, February 2009. ISBN 978-1-58603-929-5.

2 Hans Kleine Büning and Oliver Kullmann. Minimal unsatisfiability and autarkies. In [1], chapter 11, pages 339–401. ISBN 978-1-58603-929-5. DOI: 10.3233/978-1-58603-929-5-339.

3.9 Open Problems in Proof Complexity (a Personal Selection)

Massimo Lauria (KTH – Stockholm, SE)

Research in proof complexity focuses on showing lower bounds for stronger and stronger proof systems. Like circuit complexity research, the former has been stuck on difficult open problems for years. The reason is that proof complexity has been invented as a tool for studying computational complexity questions like NP vs coNP.

We propose open problems with a very different motivation. We think that proof complexity approach is useful in an algorithmic setting like combinatorial optimization.

We discuss the problem of finding a k-clique in a graph, using a SAT solver. For $k \ll n$ (as in parameterized complexity theory) we still do not know if there is something better than brute force search to prove that such clique do not exists.

We also discuss the relation between proof systems and approximation. It is known that many approximation algorithms can be proved to be correct in some geometric proof systems: this implies unconditional inapproximability results. While all known lower bounds are rank based, it is open if it is possible to lower bound the length of proofs (i.e. the running time of the algorithms).

3.10 Complexity Classifications for Different Satisfiability Problems

Arne Meier (Leibniz Universität Hannover, DE)

In this talk we introduce the audience to the techniques around Post's lattice [1]. Hereby we define the notion of Boolean clones in terms of a closure operator in means of superposition (introduction of fictive variables, arbitrary composition) applied to a finite set of Boolean functions. We explain how the lattice enables us to state complexity classifications of Boolean problems in a very structured and complete way.

The central motivation of this approach is to understand the inherent structure of a given Boolean problem and the possible connection of the underlying difficulty to a specific set of Boolean functions or, in fact, to the presence of a single Boolean function.

Next we describe several general steps which can be done always when one works with Post's lattice and demonstrate the power of this tool in following Lewis classification of propositional SAT from 1979 in complete detail [2].

Finally we visit the temporal logic CTL (Computation Tree Logic) and give an intuition about how the complexity landscape looks for this logic satisfiability problem which refers to the results in [3].

References

1 E. Post. *The two-valued iterative systems of mathematical logic.* Annals of Mathematical Studies, Vol. 5, pp. 1–122, 1941.

2 H. Lewis. *Satisfiability problems for propositional calculi.* Math. Sys. Theory, Vol. 13, pp. 45–53, 1979.

3 A. Meier and M. Mundhenk and M. Thomas and H. Vollmer. *The Complexity of Satisfiability for Fragments of CTL and CTL^**. International Journal of Foundations of Computer Science, No. 5, Vol. 20, pp. 901–918, 2009.

3.11 A Fragment of Dependence Logic Capturing Polynomial Time

Julian-Steffen Müller (Leibniz Universität Hannover, DE)

Joint work of Ebbing, Johannes; Kontinen, Juha; Müller, Julian-Steffen; Vollmer, Heribert
Main reference J. Ebbing, J. Kontinen, J.-S. Müller, H. Vollmer, "A Fragment of Dependence Logic Capturing Polynomial Time," arXiv:1210.3321v2 [cs.LO].
URL http://arxiv.org/abs/1210.3321

In this talk we study the expressive power of Horn-formulae in dependence logic and show that they can express NP-complete problems. Therefore we define an even smaller fragment D-Horn* and show that over finite successor structures it captures the complexity class P of all sets decidable in polynomial time.

3.12 Relating Proof Complexity Measures and Practical Hardness of SAT

Jakob Nordström (KTH – Stockholm, SE)

Joint work of Järvisalo, Matti; Matsliah, Arie; Nordström, Jakob; Živný, Stanislav
Main reference M. Järvisalo, A. Matsliah, J. Nordström, S. Zivný, "Relating Proof Complexity Measures and Practical Hardness of SAT," in Proc. of the 18th Int'l Conf. on Principles and Practice of Constraint Programming (CP '12), LNCS, Vol. 7514, pp. 316–331, 2012.
URL http://dx.doi.org/10.1007/978-3-642-33558-7_25

Boolean satisfiability (SAT) solvers have improved enormously in performance over the last 10-15 years and are today an indispensable tool for solving a wide range of computational problems. However, our understanding of what makes SAT instances hard or easy in practice is still quite limited. A recent line of research in proof complexity has studied theoretical complexity measures such as length, width, and space in resolution, which is a proof system closely related to state-of-the-art conflict-driven clause learning (CDCL) SAT solvers. Although it seems like a natural question whether these complexity measures could be relevant for understanding the practical hardness of SAT instances, to date there has been very limited research on such possible connections.

This work sets out on a systematic study of the interconnections between theoretical complexity and practical SAT solver performance. Our main focus is on space complexity in resolution, and we report results from extensive experiments aimed at understanding to what extent this measure is correlated with hardness in practice. Our conclusion from the empirical data is that the resolution space complexity of a formula would seem to be a more fine-grained indicator of whether the formula is hard or easy than the length or width needed in a resolution proof. On the theory side, we prove a separation of general and tree-like resolution space, where the latter has been proposed before as a measure of practical hardness, and also show connections between resolution space and backdoor sets.

3.13 Exponential Lower Bounds for the PPSZ k-SAT Algorithm

Dominik Scheder (Aarhus University, DK)

Joint work of Chen Shiteng; Scheder, Dominik; Talebanfard, Navid; Tang, Bangsheng
Main reference S. Chen, D. Scheder, N. Talebanfard, B. Tang, "Expoential Lower Bounds for the PPSZ k-SAT Algorithm," to appear at SODA 2013.
URL http://users-cs.au.dk/dscheder/dscheder-homepage/publications/ChenSchederTalebanfardTang.pdf

In 1998, Paturi, Pudlak, Saks, and Zane presented PPSZ, an elegant randomized algorithm for k-SAT. Fourteen years on, this algorithm is still the fastest known worst-case algorithm. They proved that its expected running time on $k-$CNF formulas with n variables is at most $2^{((1-\epsilon_k)n)}$, where $\epsilon_k = \Omega(1/k)$. So far, no exponential lower bounds at all have been known.

In this paper, we construct hard instances for PPSZ. That is, we construct satisfiable k-CNF formulas over n variables on which the expected running time is at least $2^{((1-\epsilon_k)n)}$, for ϵ_k in $O(log^2(k)/k)$.

3.14 Stochastic Local Search for SAT and Make versus Break

Uwe Schöning (Universität Ulm, DE)

Joint work of Balint, Adrian; Schöning, Uwe
Main reference A. Balint, U. Schöning, "Choosing Probability Distributions for Stochastic Local Search and the Role of Make versus Break," in Proc. of the 15th Int'l Conf. on Theory and Applications of Satisfiability Testing (SAT'12), LNCS, Vol. 7317, pp. 16–29, 2012.
URL http://dx.doi.org/10.1007/978-3-642-31612-8_3

Given an assignment a to a CNF formula and a Boolean variable x, MAKE=MAKE(a, x) is the number of clauses which go from false to true when flipping x's assignment. BREAK= BREAK(a, x) is the number of clauses which go from true to false when flipping x.

It seems natural that MAKE – BREAK is a natural measure to base decisions about selection of flipping variables about it. Our experiments show that BREAK is the more important parameter, actually MAKE can be ignored totally - as long as the flip variable x stems from a clause which is false under the actual assignment a.

Another experimental observation is that a 3SAT algorithm based on flipping probabilities (For those variables x as mentioned above) which are proportional to $(1+\text{BREAK}(a, x))^{-3}$ works very well.

3.15 A Satisfiability-Based Approach for Generalized Tanglegrams on Level Graphs

Ewald Speckenmeyer (Universität Köln, DE)

Joint work of Wotzlaw, Andreas; Speckenmeyer, Ewald; Porschen, Stefan
Main reference A. Wotzlaw, E.Speckenmeyer, S. Porschen, "Generalized k-ary tanglegrams on level graphs: A satisfiability-based approach and its evaluation," Discrete Appl. Math., pp. 2349–2363, Volume 160, Issues 16–17, November 2012.
URL http://dx.doi.org/10.1016/j.dam.2012.05.028

A tanglegram is a pair of (not necessarily binary) trees on the same set of leaves with matching leaves in the two trees joined by an edge. Tanglegrams are widely used in computational biology to compare evolutionary histories of species. In this work we present a formulation of two related combinatorial embedding problems concerning tanglegrams in terms of CNF-formulas. The first problem is known as the planar embedding and the second as the crossing minimization problem. We show that our satisfiability-based encoding of these problems can handle a much more general case with more than two, not necessarily binary or complete, trees defined on arbitrary sets of leaves and allowed to vary their layouts. Furthermore, we present an experimental comparison of our technique and several known heuristics for solving generalized binary tanglegrams, showing its competitive performance and efficiency and thus proving its practical usability.

(Slides: http://e-archive.informatik.uni-koeln.de/id/eprint/693)

3.16 State-based Satisfiability

Sean Weaver (University of Cincinnati, US)

State-based Satisfiability (SAT),a variant of SAT that uses state machines to represent constraints. Using this constraint representation allows for compact representations of SAT problem instances that retain more ungarbled user-domain information than other more common representations such as Conjunctive Normal Form. State-base SAT also supports earlier inference deduction during search, the use of powerful search heuristics, and the integration of special purpose constraints and solvers.

3.17 Finiteness Conjecture on Hitting Unsatisfiable Formulas

Xishun Zhao (Sun Yat-sen University – Guangzhou, CN)

Joint work of Zhao, Xishun; Kullmann, Oliver

In this talk we propose the following so-called finiteness conjecture on Hitting unsatisfiable formulas: For every k, there is a number n_k such that every hitting unsatisfiable formula with deficiency k has at most n_k propositional variables if in the formula every literal occurs at least twice. Here, deficiency of a CNF formula is the difference between the number of clauses and the number of variables. Please note that unsatisfiable hitting formulas must be minimal unsatisfiable (MU), that is, deleting any clause results in a satisfiable formula. From some known results on MU, we can see that the conjecture holds for $k \in \{1, 2\}$. In this talk a proof of the conjecture for $k = 3$ is presented. In the proof the singular DP (for short sDP) reduction plays an important role. sDP reduction is Davis-Putnam resolution applied on a variable which or whose negation occurs only once. Another proof trick is the splitting. If F is a hitting unsatisfiable and x is a variable occurring in F, then $F[x = 1]$ and $F[x = 0]$ are also hitting unsatisfiable.

Participants

- Olaf Beyersdorff
University of Leeds, GB
- Uwe Bubeck
Universität Paderborn, DE
- Catarina Carvalho
University of Hertfordshire, GB
- Nadia Creignou
Université de Marseille, FR
- Stefan Dantchev
University of Durham, GB
- Evgeny Dantsin
Roosevelt Univ. – Chicago, US
- Anuj Dawar
University of Cambridge, GB
- Arnaud Durand
University Paris-Diderot, FR
- Johannes Ebbing
Leibniz Univ. Hannover, DE
- Uwe Egly
TU Wien, AT
- John Franco
University of Cincinnati, US
- Nicola Galesi
University of Rome “La Sapienza”, IT
- Heidi Gebauer
ETH Zürich, CH
- Andreas Goerdt
TU Chemnitz, DE
- Miki Hermann
Ecole Polytechnique – Palaiseau, FR
- Timon Hertli
ETH Zürich, CH
- Edward A. Hirsch
Steklov Institute – St. Petersburg, RU
- Kazuo Iwama
Kyoto University, JP
- Jan Johannsen
LMU München, DE
- Lefteris M. Kirousis
University of Athens & Computer Technology Institute & Press “Eudoxus”
- Hans Kleine Büning
Universität Paderborn, DE
- Donald Ervin Knuth
Stanford University, US
- Juha Kontinen
University of Helsinki, FI
- Alexander S. Kulikov
Steklov Institute – St. Petersburg, RU
- Oliver Kullmann
Swansea University, GB
- Massimo Lauria
KTH – Stockholm, SE
- Victor W. Marek
University of Kentucky, US
- Barnaby Martin
Middlesex University, GB
- Arne Meier
Leibniz Univ. Hannover, DE
- Julian-Steffen Müller
Leibniz Univ. Hannover, DE
- Jakob Nordström
KTH – Stockholm, SE
- Ramamohan Paturi
University of California – San Diego, US
- Rahul Santhanam
University of Edinburgh, GB
- Dominik Scheder
Aarhus University, DK
- Henning Schnoor
Universität Kiel, DE
- Uwe Schöning
Universität Ulm, DE
- Martina Seidl
University of Linz, AT
- Robert H. Sloan
Univ. of Illinois – Chicago, US
- Ewald Speckenmeyer
Universität Köln, DE
- Stefan Szeider
Vienna Univ. of Technology, AT
- Jacobo Toran
Universität Ulm, DE
- Heribert Vollmer
Leibniz Univ. Hannover, DE
- Sean Weaver
University of Cincinnati, US
- Xishun Zhao
Sun Yat-sen University – Guangzhou, CN

Report from Dagstuhl Seminar 12472

Is the Future of Preservation Cloudy?

Edited by

Erik Elmroth[1], Michael Factor[2], Ethan Miller[3], and Margo Seltzer[4]

1 **University of Umeå, SE,** `elmroth@cs.umu.se`
2 **IBM Research – Haifa, IL,** `factor@il.ibm.com`
3 **University of California – Santa Cruz, US,** `elm@cs.ucsc.edu`
4 **Harvard University, US,** `margo@eecs.harvard.edu`

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 12472 "Is the Future of Preservation Cloudy?". Our seminar was composed of a series of panels structured as a series of brief presentations followed by an open discussion. The seminar started with a session introducing key concepts and definitions and illuminating the vast array of perspectives from which attendees were addressing issues of cloud and preservation. We them proceeded into a discussion of requirements from different types of communities and a subsequent discussion on how to protect the data and ensure its integrity and reliability. We next considered issues related to cloud infrastructure, in particular related to management of the bits and logical obsolescence. We also considered the economics of preservation and the ability to reuse knowledge. In addition to these pre-planned panels, we had three breakout sessions that were identified by the participants: automated appraisal, design for forgetting, and PaaS/SaaS for data preservation. After the executive summary, we present summaries of the panels and reports on the breakout sessions, followed by brief abstracts from a majority of the seminar participants describing the material they presented in the panels.

Seminar 19.–21. November, 2012 – www.dagstuhl.de/12472

1998 ACM Subject Classification D.4.3 File Systems Management, D.4.5 Reliability, D.4.6 Security and Protection, E.2 Data Storage Representations, E.5 Files, H.3.2 Information Storage, H.3.4 Systems and Software, H.3.6 Library Automation, H.3.7 Digital Libraries, K.6.5 Security and Protection

Keywords and phrases long-term preservation, cloud storage, provenance, obsolescence, data access, storage systems

Digital Object Identifier 10.4230/DagRep.2.11.102

1 Executive Summary

Erik Elmroth
Michael Factor
Ethan Miller
Margo Seltzer

Two significant trends in data management are emerging: data is moving to cloud infrastructures and an increasing fraction of data produced is born digital. We risk losing all record of born digital data if we do not take explicit steps to ensure its longevity. While each of these trends raises its own set of questions, our seminar began with two fundamental questions at the intersection of these trends: What role should the cloud play in preservation? What steps should we be taking now to preserve the future of today's digital artifacts?

We addressed these two questions by bringing together a diverse cohort of approximately thirty participants. Our participants consisted of researchers from both academia and industry, representatives from cloud providers, and archivists and librarians from memory institutions. Every participant was responsible for some aspect of the program, and the workshop was characterized by lively debate. There were four primary outcomes of the workshop:

1. We identified key functional requirements that are critical if cloud infrastructures are to be used for long-term digital preservation.
2. We identified topics where we were unable to reach agreement; since we are trying to look into the future, while not satisfactory, it seems likely we will need to wait until the future to resolve these debates.
3. We identified several specific problems requiring further work and brought together groups of people interested in pursuing those areas.
4. We identified several areas that we were not able to address, either because we lacked the expertise in the room or we ran out of time; these areas represent opportunities for subsequent workshops.

Perhaps the most pressing issue with respect to existing cloud infrastructures is the lack of standardized APIs. If data are to outlive any particular organization, then it is crucial that archives span organizational boundaries; standardized APIs make this dramatically easier and more robust. There was also agreement that some form of automated appraisal was important, but there were no concrete ideas about how to do it.

We had lively debate around the long term cost of cloud storage, in particular public clouds; since this debate depended upon assumptions of future costs, the future will ultimately resolve the debate. We also had much discussion around the importance of logical preservation and whether the modern world, with readily available open source viewers has made the need for logical preservation obsolete.

Several small working groups coalesced around the areas of: archival exit (how do you get data out of an archive), the technical design of preservation-as-a-service (PaaS), technologies for ensuring that data is "forgotten", and searching distributed archives. We are hoping to see these small groups evolve into productive collaborations that continue the work begun at the seminar.

Finally, there were a number of areas related to using the cloud as a preservation service that we were unable to address. For example, what legal issues arise if companies undertake digital archival initiatives? Is there a legal definition of "deletion" of data, and is it practical? Where does "record management" end and "archival" begin? Who is the customer for long term preservation? Is it the data provider? Or perhaps it's the data consumer? What happens to archived data if payment cannot be made? What is the economic model behind long term archival? These and other questions provide ample opportunity for further workshops on this topic.

Organization

The workshop was organized around a series of 90-minute sessions, each of which began with one or more short presentations followed by a moderated discussion. We had one person scribe each session and the session moderators produced the session summaries that appear in this report documenting each session. We also devoted one session to smaller breakout groups, who reported back in our closing session.

2 Table of Contents

Executive Summary

Panel Discussions and Session Summaries

Working Groups

Overview of Talks

3 Panel Discussions and Session Summaries

3.1 Opening Session

Mary Baker (HP Labs – Palo Alto, US)

One of the lovely and remarkable features of this Dagstuhl seminar was the diversity of disciplines represented by the participants. Participants introduced themselves as coming from a wide variety of scientific, industrial, government, cultural, and academic institutions. Their areas of expertise included digital preservation technologies, storage and database products, huge cloud storage applications, memory institutions, digital curation, provenance of digital content, medical records and their associated policies, trace archives of distributed systems, the economics of digital preservation, scientific computing, supercomputing, and so forth. To illuminate the different issues faced by participants, we asked everyone to describe what digital preservation means to their communities and what preservation problems the cloud solves and does not solve for them.

This diversity of participants also posed a challenge for us: finding a common vocabulary and set of concepts for digital preservation so we could avoid confusion and make forward progress. We therefore used the first session to address the problem. We introduced the goals of digital preservation by claiming that digital assets stored now should remain accessible, usable and undamaged for as long as desired – beyond the lifetime of any particular storage system, storage technology, or storage vendor, and that this must be done affordably. The main discussion of these goals centered around the meaning of undamaged, since it has different meanings depending on the kind of asset being preserved and the preservation purpose. For some digital assets undamaged means the bits must not change. For others the bits may change but the meaning must remain the same. For yet other assets, the contents need to remain usable.

The terminology and concepts we presented included:

- “physical” or “ bit preservation” (and why it is still a challenge to do affordably),
- “logical preservation” (and the Performance Model used by the National Archives of Australia to illustrate the problem),
- “metadata” (which is too broad a term according to some of the participants), and
- what different communities mean by “preservation.”

When we preserve an asset – what are we preserving? For instance, for a book do we just save the text? How about images of the pages? What about saving the political and cultural context in which it was published? For applications, do we save the entire ecosystem in which they run? Or are screen shots of the various user interface activities sufficient?

We deliberately avoided defining “the cloud” and this came back to haunt us later in the seminar!

3.2 Domain Specific Needs

Erik Elmroth (University of Umeå, SE)

This panel, focusing on domain specific preservation needs, included Ian F. Adams (University of California, Santa Cruz, CA, USA), Dirk Nitschke (Oracle, Hamburg, Germany), and Gillian Oliver Victoria University of Wellington, New Zealand) and was moderated by Erik Elmroth (Umeå University, Sweden).

The goal of our panel was to provide a concrete examples, from a variety of domain, about what types of information need to be preserved in the cloud, for how long must they be preserved, for whom, by whom and in what type of cloud? We wanted to move beyond common preservation requirements and focus on requirements specific to one or more domains. For example, when discussing memory institutions, forgetting became important; it is equally important to intentionally and thoughtfully decide to not remember (preserve) things as it is to select things to be remembered.

A second topic that created much lively discussion concerned the risks of storing information in public clouds. This flowed seamlessly into a discussion of existing laws and the challenges of jurisdiction – how do cloud providers and customers come together when legal requirements differ for each of them? Someone described the Megaupload case where customer content was seized, because some customers had uploaded data without appropriate copyrights, as an example of the legal complexities that arise.

The legal discussion then flowed naturally into a discussion concerning the difference between selective archival and censorship and whether and whether it is at all feasible for archivists to decide what, from the massive data stored in clouds, should be preserved. For example, the Internet archive uses a statistical, or perhaps random, approach to archiving, not a human-centered manual one.

When the conversation moved to scientific data archiving and high-performance computing applications with very large data sets from climate, particle colliders, etc., we began trying to distinguish between data and information. In cases where data can be reproduced (which was agreed to be a fundamental concept), should archives leverage this capability to reduce capacity needs. For example, if a person's DNA sequence has been computed, need we save that sequence or can we discard it and then resurgence it later? Of course, once we begin discussing data reproduction, the challenge of preserving software and execution environments becomes critical.

When taking the business perspective, the discussion touched upon cases where data are truly mission critical and data loss must be avoided "at all costs." In fact, the requirements may be even more stringent – in some cases, it is not only necessary to make it possible to obtain archived data, it might need to be always available relatively quickly. This generates both bandwidth and latency requirements. As with memory institutions, the issue of forgetting also came up in the business context. For example, regulations may require retaining financial data for a defined period, but beyond that period the data can become a liability. Unlike memory institutions, businesses do not have archivists, so end-users, untrained in archival and digital preservation, are responsible for identifying data to archive. A long discussion on incentives led to the conclusion that finding a way to make money out of archived data was probably a more effective incentive than a legal frameworks.

After we covered the domains individually and had a more encompassing discussion, someone observed that there is no accidental digital preservation, while there is accidental physical preservation. Such accidental physical preservation, e.g., finding a cache of lost letters in an attic, has turned out to be crucial for historical investigations. The discussion on whether we can ensure that something is not preserved led to our distinguishing between “digital preservation” and “digital archaeology”, e.g., the ability to recreate old computer games when the physical media has become obsolete.

3.3 Protecting the Data

Margo Seltzer (Harvard University, US)

Our panel, comprised of Jean Bacon (University of Cambridge), Ewnetu Bayuh Lakew (Umea University), Peter Pietzuch (Imperial College London), and Ken Moody (University of Cambridge), and moderated by Margo Seltzer (Harvard University) represented a diverse set of opinions on what protection meant, how it applied to preservation, and what about the cloud made it different. The presentations addressed issues ranging from the legal challenge when data moves across geopolitical borders, to the privacy challenge when the data being stored was health-related, to the technical challenges of describing and translating security policies.

There were two distinct parts of the session – first there was discussion about some of the key aspects of protecting data and second, it became clear that the attendees needed to come together to agree upon terminology surrounding clouds. After much discussion on this second point, Alexandr Iosup (Delft) provided the NIST definition of a cloud, “cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.” This provided a specific definition that was used for the rest of the workshop.

The technical discussion ranged over three main topics: the relationship between provenance and preservation, the role of the cloud in preservation, and the differences in preserving private versus public data. Some participants questioned what provenance had to do with preservation, but there was consensus that documenting ingest processes, migration efforts, and emulation requirements was critical to ensuring meaningful access to digital objects in the long term.

Unsurprisingly, the group spent a great deal of time discussing the role of the cloud in preservation – asking and discussing questions such as whether the cloud is simply a technology, a delivery mechanism, a platform, a business model, etc. Today, clouds are clearly only part of a solution – they are not fundamentally preservation systems. However, if they are to be part of a long term preservation solution then there are important avenues for research and development. For example, while many existing infrastructures claim to support “standard” APIs (e.g., Amazon S3), experience with LOCKSS and other systems suggests that they are not; you cannot write an application or service to a single API and have it run across multiple providers. Therefore, vendor lock-in, is a huge issue when contemplating long term preservation. More fundamentally, the group asked whether commercial providers were the right parties to provide preservation platforms, or whether governments or memory institutions were better suited to the task. No clear consensus emerged from this part of the discussion.

Finally, the group turned to questions surrounding the preservation of both public and private data. Private data requires adequate security, but what can be considered adequate if we are trying to preserve objects for tens or even hundreds of years? With today's computational resources, it's possible to break a lot of encryption with a hundred years of compute time; with future breakthroughs, what can we consider truly protected? In many ways, public data presents a much easier problem, because archives need only concern themselves with ensuring authenticity. However, it was pointed out that some public data starts out as private data, e.g., census data. Therefore, it is not sufficient to simply focus on public data and ignore the challenge surrounding private data. There was no group agreement on what kinds of guarantees for long term security were sufficient.

3.4 Reliability and Integrity

Lawrence You (Google Inc, US)

The panel on Reliability and Integrity had four presenters: Gerhard Schneider, University of Freiburg, Germany; Nikos Chondros, University of Athens, Greece; André Brinkmann, University of Mainz, Germany; and Lawrence You, Google Inc, USA.

Our four panelists each made short presentations on different facets on cloud storage reliability and integrity, approached as a user (Schneider), verifier (Chondros), exploiting provider diversity (Brinkmann), and as a provider (You).

Common themes were that cloud services vary, due to different product offerings and commitments. There was consensus during the presentations and questions that data preservationists must use multiple cloud providers to ensure long-term integrity and reliability if they are to use them at all. Presenters raised many points on numerous challenges that still exist with cloud storage services: that storage cloud services alone are insufficient for integrity when access also requires software; that verification requires access and re-fetching copies, and costs must be factored in; that cheap storage is unreliable; and that consumer versus enterprise storage pricing meet different business needs.

The discussion following the presentations produced a lot of questions, particularly in regard to measurement of reliability and types of failures to guard against. Cloud as a model for preserving data has some requirements that differ due to time scales that are long, and reliability requirements where failure modes of error rates (for example, the number of "9"s of reliability) are independent of trust in the cloud providers, which are businesses. Diversity is important, but so is cost of the service and cost of failure.

We wrapped up the discussion with the basic question:

Do we think reliability and integrity are a solved problem?

There was some disagreement, but there were a number of questions/comments from the wrap-up:

We have (meaning people here use) a solution using a combination of techniques. It's not possible to say it's solved, because not all information is classified, or it is classified incorrectly. Different cloud pricing and business models differentiate service/reliability. Cloud providers have a core competency in bit preservation, while archivists do not. Will the cloud cost too much? What is the real cost of reliability (for preservationists)? Including cost of failure?

Failure comes in many forms: natural disasters, insider abuse, black swans. A moral hazard: cloud providers want to claim 11 nines and the buyer wants to believe such claims, but will everyone be gone when the claims are tested? Is reduced reliability and more providers a better solution?

3.5 Preservation Storage / Cloud Services Issues

Hillel Kolodner (IBM Research – Haifa, IL)

The panel had four presenters: Joanne Syben (Google) (lead), Alexandru Iosup (Delft University of Technology), Sam Fineberg (HP), and Hillel Kolodner (IBM).

Joanne Syben spoke about technical aspects of current commercial storage clouds and issues regarding user-owned and user-derived data. In current storage clouds, there are several interesting combinations of durability and availability. For example, Amazon currently offers Reduced Redundancy Storage for low durability and high availability, Standard S3 for high durability and high availability, and Glacier for high durability and low availability. Joanne mentioned life cycle issues that need to be automated: migration to new hardware, migration to cheaper hardware for cold data, and deletion (including high volume deletion). She also discussed the appropriateness of some hardware options including tape and shingled disks.

Joanne also discussed the distinction between user-owned and user-derived data and the resulting technical and legal issues. A significant technical issue around user-owned data is how should it be curated? Methods include relevance ranking, time based ranking and explicit annotations. Regarding legal issues, given that data can be considered an inheritable asset, what rights should heirs have to curate it? What about jointly owned data in the event of a divorce? Search quality depends on user-derived data. User-derived data also needs to be curated, yet it is hard to predict what user-derived data will be useful in the future. There are also privacy issues.

Following Joanne's presentation there was a discussion about the difficulty of deleting data; this discussion led to one of our breakout sessions. There are both legal and technical issues. Legally there are regulations, e.g., a legal definition of secure deletion. Technically, there may be many copies of a data item that need to be deleted, e.g., multiple copies on-line and multiple copies in backups, and it might not be possible to find all of the copies.

Alexandru Iosup presented the idea of a distributed systems memex, i.e., logging and preserving the entire history of a distributed system; he also presented several experiments that he and his team have done in this regard. These include the Grid Workloads Archive, containing six online traces, the Failure Trace Archive, containing 25 online traces, and the Game Trace Archive. Alexandru went on to discuss the benefits of using clouds for the storage of the traces, e.g., simplification of management, reduced cost for the infrastructure, and the availability of computation close to the data. However, there are also challenges: processing close to the data leads to vendor lock-in, and the challenges of migration and security. Finally, Alexandru presented several experiments that he and his team have done regarding the performance and performance variability of AWS services (EC2 and S3) and DropBox. This presentation was followed by a discussion on long term funding for archives and curation of archives; although, no conclusions were reached.

Sam Fineberg presented some advantages and drawbacks of using clouds for preservation, and then briefly introduced SIRF (Self-contained Information Retention Format). Advantages of using cloud storage for preservation include the simplification of migration and economy of scale. Drawbacks include the elimination of transparency and control. Sam argued that preservation should be considered SaaS (software as a service) rather than IaaS (infrastructure as a service).

There was a discussion about what would be the SaaS services; possibilities include logical migration, physical migration, and verification services. David R. pointed out that there are already preservation-as-a-service offerings available today, but why should we trust such a service. One service called out in particular was DuraCloud which has the benefit of being open source. There was also a discussion on whether we should be striping the preserved data over multiple such services.

Sam also presented SIRF, which is Storage Networking Industry Association effort to define the logical equivalent to storage boxes which can be seen in a physical archive. It needs to be self-describing, self-contained, and extensible. One concern that was raised was the scoping of metadata to a logical container in SIRF; is it too easy with such scoping to have broken links, e.g., if single data set doesn't map cleanly to a container.

Hillel Kolodner raised issues regarding the requirements from a storage cloud to support a preservation system. For example, should a preservation system provide deep support for preservation, e.g., very high reliability and end-to-end security. Or should storage clouds provide basic support (e.g., lower levels of reliability) and preservation systems ensure high reliability themselves, e.g., by storing data on multiple clouds. In this case, what is the basic support that is needed in each cloud and how should the preservation system leverage it? Issues include the costs for replication, integrity checking and provenance. For example, we would not want three way replication per cloud when replicating across three clouds – this would mean keeping nine copies. Integrity checking would also be occurring repeatedly in each of the clouds and then also by the preservation system. Provenance would also be hard because it would be necessary to do full provenance on each cloud and also in the preservation system.

Hillel also raised the question whether there are advanced features that storage clouds could provide that facilitate preservation systems. For example VISION Cloud provides support for rich metadata and allows objects to be found based on their metadata values. It also supports safe and secure computation in the storage system, e.g., which can be used for integrity checking. Finally, storage clouds can provide support for the secure handling of data, e.g., secure isolation between the data of tenants and users, geographic constraints on the placement of data and secure delete.

Following Hillel's presentation there was a discussion about some of the issues raised. A cloud provider can really go out of business. One solution that was suggested is that the data owners can physically own their own disks. However, this raises several issues. It is less elastic. It is harder to achieve economies of scale. And it could raise problems when necessary to migrate data to new physical media.

3.6 Handling Logical Obsolescence

Liuba Shrira (Brandeis Univ. Waltham, US)

The panel had four presenters, Michael Factor (IBM), Natasa Milic-Frayling (Microsoft), Matthias Grawinkel (Universität Mainz), and Liuba Shrira (Brandeis) (lead).

Matthias Grawinkel surveyed the factors governing the longevity of archived objects, and considered the constraints on the storage media imposed by different longevity time scales. He then discussed the importance of techniques for preserving the object interpretation environment, describing the extreme case of the strong versioning approach adopted by NASA, which allows reproduction of the exact processing environment.

Natasa Milic-Frayling focused on the computational nature of the digital artifacts. She defined digital preservation as "enabling digital artifacts to be instantiated in a contemporary environment," and argued that the key to ensuring long-term preservation is providing efficient software development environments for developing "bridging components" such as format translators and virtual machine adapters. She then described an Azure based service for format migration designed using this approach, as part of the SCAPE project, that is extensible in both formats and data storage.

Michael Factor shared lessons from his work on the EU funded ENSURE project, focusing on three points,

1. supporting requirements-based preservation plans that protect different data in different ways;
2. preservation of metadata, including ensuring that meaningful cost effective metadata is available for all entities at ingest, since it may be impossible to reconstruct metadata afterwards;
3. automation of transformation and verification, achieved in ENSURE by a combination of a workflow engine, virtual appliances for short term preservation, and computational storage supporting transformation and verification for long term preservation.

Liuba Shrira described an efficient just-in-time transformation service for handling logical obsolescence in a cloud-based preservation system. Such a migration service defers transformation until the object is used, avoiding transforming work when a new object format becomes available. The challenge for just-in-time migration is how to avoid introducing transformer dependencies on future versions. Shrira described a framework, developed in her research on database upgrades, that supports efficient just-in-time transformation and eliminates transformer dependencies on all but a single predecessor version. She raised the question of what would it take to achieve analogous transformation system properties at different levels in the cloud-based preservation system software stack.

The follow-up discussion centered around two key issues, the implications of managing logical obsolescence at different levels in the software/hardware stack and the challenges that arise when the digital artifacts to be preserved and the techniques of preserving them have a computational component.

The discussion started by reviewing the accepted roles of migration and emulation in handling logical obsolescence, (migration being the standard tool with a downside of losing information, and emulation being the fall-back for when migration fails), and then discussing the open problems with both approaches.

An issue for emulation is at what level to emulate. VMs are a standard emulation level, because they represent a "slow moving layer". This works well in the short term but poses difficulties for long term preservation. VMs have not been designed with migration in mind; they are complex and have dependencies that need to be handled when eventually migration takes place. Could a different layer work better? An interesting example of a different layer is IBM's successful move to a different hardware architecture (S400) at an internal OS level rather than machine level. A standardized API for emulation would have a dramatic impact, but few participants were optimistic this would happen any time soon.

An issue for migration is that the layer above may become obsolete as well, making it hard to resort to emulation at the point where migration does not work anymore. For example, it is hard to rebuild renderers after they are gone without a good enough "live" artifact that preserves the experience. Of course, emulation does not have to be complete, but it is hard to know what path will be useful in the future, especially, if we want to ask old data new questions, in addition to old questions.

Another issue for migration is that transformers are programs too and need an environment in which to run. Could the cloud help with keeping transformers alive by maintaining libraries?

Logical obsolescence of digital artifacts manifested (and consumed) using computation rises several issues. On the one hand, programs may have assertions, and executable specifications, and these can be used to assemble a testable verification, e.g., to verify whether rendering is correct. A cloud may even help with high-fidelity recursive emulation at multiple levels that results in substantial computational overhead. On the other hand, the more complex the digital experience, the harder we may need to think what obsolescence might mean. For example, the gaming industry has economic incentives in preserving gaming experiences, re-issuing them in the future, just like Hollywood is preserving old movies. What does it mean to preserve a multi-player game experience? Similarly, what does it mean to preserve a spreadsheet combining live results from multiple distributed data streams? The participants agreed that a new principled approach must be developed for preservation of "digital experiences".

Throughout the discussion the participants considered how cloud infrastructure could impact the handling of logical obsolescence, we had more questions than answers, but the consensus was that by further separating the owner of a digital artifact from the infrastructure that handles obsolescence, the cloud makes it harder to exploit native approaches. On the other hand, the cloud could provide the advantage of more plentiful computational resources or have available a wider choice of preserved standard environments.

3.7 Knowledge Re-use

David Giaretta (APA, GB)

URL https://dl.dropbox.com/u/6959356/Knowledge use notes.docx

This session had three presenters: David Giaretta (APA), Christoph Becker (TU Wien), and Milena Dobreva (University of Malta). It covered the importance of knowledge re-use in preservation.

The session began with David Giaretta's talk "Knowledge, Value and Services for preservation with some thoughts on clouds". Drawing on quotes from Neelie Kroes and a report from the High-Level Group on Scientific Data, David illustrates the desire that results of publicly funded research provide valuable assets and claims that preservation, by its nature of facilitating reuse, makes assets more valuable. He goes on to claim that while traditional archival focuses on documents, we should focus more on data, because data is more challenging – data value increases when we know the semantics of the data and can combine it with other, potentially massive, pieces of data.

This leads to the challenge David puts before us: how do we make the stuff we preserve more valuable? David proposed that a way to look at what is needed for preservation was to look at threats arising from changes in technology (hardware and software), environment, e.g. e.g. name resolvers, and tacit knowledge of users

To counter these threats the mantra one tends to hear from libraries is: "Emulate or migrate." However, in line with the argument above one can see that emulation works well with data only in special cases, because one can repeat what was done in the past rather than doing new things.

Turning to what kinds of semantics – or, more generally, knowledge – is required, the simplest are things such as units of measure, etc. Next comes "Representation Information," the OAIS term for what many might call metadata (a term David chastises us all for using, because it is ill-defined). Noting that emulators are also a type of Representation Information one can restate the mantra as "Add Representation Information or Transform. Or move to another repository". Another complexity which one must deal with when one wishes to keep data valuable by keeping it usable is to recognise that any piece of Representation Information must itself be usable. This introduces a potentially problematical recursion, which OAIS resolves through the concept of a "Designated Community." The advantage is that this requires ways to figure out how much Representation Information needs to be provided for some community to be able to understand and use the data. This is precisely what is needed to add value to data by making it more widely usable and for data from many sources to be combined easily.

The CASPAR and its successor the SCIDIP-ES projects are addressing the preservation and use of digitally encoded information by providing tools and services to supplement what a repository does by allowing curators to

1. know something has changed
2. identify the implications of that change
3. decide on the best course of action for preservation
4. determine what RepInfo we need to fill the gaps

Michael Factor asked if there is added value by putting the services in the cloud, to which David's reply was that this must be the case – allowing greater resilience and pooling of resources.

Liuba asked for an example of adding value by combining data. David pointed to all the climate change work, the studies combining sociological data, health data with satellite information about temperature changes. Christoph added examples such as combining 60 years of nutritional data with climate change.

The second talk was by Christoph Becker, titled "Some thoughts on clouds, preservation, and knowledge". Christoph focused on what he called Creative Friction between the important factors that interact. Looking in detail at preservation one needed to address bitstream preservation as well as, for example, the logical layer, semantics, costs, etc.

A useful source of information on this is Open Research Challenges in DP, wiki at http://sokrates.ifs.tuwien.ac.at. Another view was of digital preservation as communication with the future, however with several potential twists: at the time of reception (1) the message may no longer exist (2) there may be no sender (3) there may be no easily available encoder to check against and (4) the recipient may not be the original addressee.

Assuming we have the bits in the future we could treat those bits as a black box, but the key question is whether we can get to the knowledge encoded in those bits. Can we do something useful with them? Can we find different knowledge in them?

We are moving towards knowledge organisations with diverse knowledge and correspondingly diverse needs of representing knowledge. Some key initial questions about organisations preserving this knowledge are: (1) what do they have (2) which capabilities does such an organisation require (3) which services do these capabilities require (4) how can both be measured and (5) what are the cost/ risk and value?

The third and final presentation was by Milena Dobreva and looked into the current context of digitisation and preservation in the memory institutions in the EU, raising the possibility that knowledge re-use may be even more cloudy than the future of digital preservation. The presentation also provided examples on capturing intermediaries' requirements for digital preservation systems.

The results of the ENUMERATE survey for 2012 into digital preservation in the EU produced some fascinating statistics and alarming conclusions. The good news is that there is a lot of activity in digital preservation:

- 83% of institutions have digital collection or currently involved in digitization
- 23% have a written DP strategy
- 33% are included in a national preservation strategy
- 30% are included in a national DP infrastructure

The bad news is that both users and curators have serious issues. Users report that "Archival practitioners" are not disciplined, the terminology they use is not consistent, hierarchical representations are unclear, the existing search tools are inadequate, and content visualization is not uniform. The curators report that they face challenges in the diversity between documents (e.g., fonts, multimedia), diversity of collections (e.g., data size for audio/video), the level of metadata (e.g., how much to include, how to collect (cost-effective), extract, select, and predict it), the absence of linguistic support in discovery and finding aids, workflows, and cost.

After these presentations, we had a lively discussion which started with Michael's question: Is there a conflict in re-use/combination (i.e. generating knowledge) vs preservation? David G had an unequivocal "no." Andre pointed out that preserving data and knowledge is just the first step and gave as an example a recent project that combined weather data with energy generator (solar, wind) data in a new context. Margo added the observation that the earlier you use the data, the more chances you have to improve the accuracy of stored data. Liuba thought that by providing the collective knowledge required for future changes, a cloud could help address the fact we don't know how data will be used in the future. Later on in the session, Ethan expanded on this point describing that the cloud can provide standard migrators.

David Rosenthal raised the concern that since 50% of the cost is paid upfront for the ingest, if one requires "extra stuff" for re-use then cost increases, but budgets are not flexible. Therefore making things re-usable may not be economically feasible. David G argued against this conclusion, because the 50% figure comes from research into rather specialised, small data sets, mostly of documents. When one deals with large volumes of data, with more uniform ingest mechanisms, then the figures will be vastly different

The key points in the remaining discussion were:

- what about preserving non-born digital artifacts?
- where are the representation information repositories and can we depend upon them?
- we need to remember access and not just collecting
- what about cost of computation versus cost of data

3.8 Economics

Ethan Miller (University of California – Santa Cruz, US)

This panel had four presenters: Ethan Miller (University of California – Santa Cruz), David S. H. Rosenthal (Stanford University Libraries), Ross King (Austrian Institute of Technology) and Raivo Ruusalepp (National Library of Estonia).

This panel discussed the economics of long-term preservation, focusing on two major issues. The first two panelists discussed the problem of forecasting the cost of long-term preservation storage and the impact of different factors on this cost. The second two panelists then discussed how these costs might be borne by users and potential funding models for long-term preservation.

The first half of the session featured two talks that described the use of economic modeling to study the costs of long-term storage, particularly as device types and costs shift and systems shift to a cloud-based model. The first panelist, David Rosenthal, examined the costs of long-term storage in the cloud, as exemplified by Amazon and others. He explained that the long-term cost of storage is dominated by the rate at which storage gets cheaper, and that this rate may be slowing. While Amazon and other large cloud providers have a large advantage in that they can "smoothe" demand for long term storage across many consumers (organizations), large cloud providers also have a captive market allowing them to somewhat artificially keep prices higher. The alternative—building a private cloud-like system—can provide a return on investment within three years, given current pricing. Ethan Miller, the next panelist, explored trends in long-term storage and discussed the interplay between changes in storage cost and desire for reliability. He noted that, as storage growth rates slow, reliability becomes increasingly important because it is more worthwhile to retain storage for longer. This has a big impact on cloud storage because it can give an advantage to devices such as solid-state storage that are currently unfeasible for long-term preservation. He suggested that architects of cloud preservation systems encourage storage designers to improve reliability, even if it means slightly higher storage costs.

The second half of the session examined the problem of economics from the preservationists' point of view. Ross King discussed the commonly-used "endowment" funding model, and noted that it was similar to a pension or Ponzi scheme in which new users pay the costs for existing data. He expected that collections would grow, allowing new data to dominate old data in size. However, since this growth may slow in the future, he suggested that we need to stop archiving everything and increasingly turn to automated appraisal to limit the amount of data we preserve. In response to a question, he added that preserved data will expand to fill available capacity, further motivating the need for automated appraisal. Other comments included suggestions to better monetize stored data, and to perhaps re-examine archived data after a period of time to see if it's still worth preserving. The final panelist, Raivo Ruusalepp, explored economics from a users' perspective, noting that digital preservation is an unfunded mandate and that few organizations even know their annual unit cost for digital preservation.

Ruusalepp then presented statistics from digital libraries, based on a study available from http://dp4lib.langzeitarchivierung.de/. This study found that staff costs are the dominant factor for the three phases of preservation: ingest, curation, and access. He then described several DCH and GRID initiatives, including a DC-Net project survey (http://www.dc-net.net) that surveyed state-of-the-art digital preservation services in 2012 (survey at http://www.dc-net.org/getFile.php?id=467). He also discussed the Indicate project (http://www.indicate-project.org) and e-Culture Science Gateway (COMETA). He concluded with a discussion of how the cloud fits into the life-cycle costs of a digital object, noting that the cloud makes it difficult to "click and forget" because of the need to monitor cloud providers and pay for storage on a per-gigabyte basis.

The session concluded with questions from other workshop attendees. The first question asked what costs went into the different phases of preservation (ingest, curation, access). David Rosenthal stated that storage costs were only relevant for a few years, and could then be ignored because of the growth in storage density. Michael Factor countered David's argument, saying that ingest, migration, and other factors make the problem worse than presented. Margo Seltzer added that cloud preservation changed the game somewhat because it allowed an organization to pay by the gigabyte-month rather than paying for an archive up front. This was followed by a discussion of who actually pays the costs for preservation services, the user who stores the data, or the one who accesses it? If they are the same person (or organization), this is a moot point, but it's often the case that they are different.

Overall, the session was successful in providing a broad overview of the economics of digital preservation in the cloud, both from the perspective of architects modeling long-term storage costs and from the perspective of users and consumers finding the funding to pay for it.

3.9 Preservation Storage & Cloud Issues

Joanne Syben (Google Inc. – Mountain View, US)

Common sense and intuition suggest that a simple approach of treating the value of data more or less equivalent to its freshness, is a good, default model for how easily accessible the data should be. Older data can automatically migrate to progressively colder and colder storage based on its age. Tradeoffs may be made between durability, availability and reliability. The current 'coldest' storage is tape, however its viability for servicing retrieval of objects from multi Petabyte or even Exabyte volumes is questionable.

The laws surrounding Data Protection Authority for many countries may confound this simple algorithm. If a user wishes to delete her account, the action needs to stretch across possibly many platforms of storage. What is the scope of the data associated with a user given the complex graphs of social media? Should the distribution of user data across different storage media be controlled by the provider or curated by the user? If curated by the user, how can this be made simple and intuitive, as well as practical? There are also many concerns about the heritability of data. Different storage media can be used to reflect the value of data, but establishing what the value is in any systematic way apart from age and possibly 'last time accessed' is an open question.

4 Working Groups

4.1 Automated Appraisal

Michael Factor (IBM Research – Haifa, IL)

The participants in the Automated Appraisal breakout session were: Ian F. Adams, André Brinkmann, Michael Factor, Ross King, Ken Moody, Gillian Oliver, and Joanne Syben.

The discussion started with level setting on the definition, the tools and the use cases. Automated appraisal refers to decisions about what to ingest and what not to ingest, priority, time to live (TTL), and semantic tagging. Tools for automated appraisal include rule-based systems, image analysis (face recognition), (near) de-duplication, natural language processing (NLP) (e.g., cross-referencing), machine learning (approximate the archivist). The use cases include regulation enforcement (e.g., health records), corporate assets, and receiving a box of "disks".

In the breakout session, the team reached the following conclusions:

- automated appraisal is hard for memory institutions, but may be easier for other domains
- it is very domain specific
- it implies new roles for curators and archivists, e.g., define rules, tweak algorithms, train machines

We concluded that a cloud can be quite beneficial for automated appraisal since it involves tasks that are bursty and computationally expensive.

In thinking about automated appraisal we turned to a discussion on the distinction between archival preservation and records management. For instance, is the storage of health records really archival or is it just record management?

Perhaps this distinction is a remnant of the paper world, but is being re-thought. It's not clear – does integrating record management and archival make things easier or harder? Which problems should be solved early in the life cycle and which problems should be solved later in the life cycle? Based upon the discussion, it seems that the answer might be domain-specific. Perhaps the answer comes down to life time? Something that has to live "long enough" requires archival "processes" while things that don't have to live "sufficiently long" perhaps don't?

In the discussion with the entire group, David R. suggested that we focus only on born-digital media, so we do not have to continually come back to the decision about the relative costs of digitization and storage. This decision on what to digitize has a significant effect on appraisal.

We then discussed the following hypothesis brought up by Michael Factor: Automated appraisal is a hard problem for memory institutions but may be much simpler for other areas (e.g., health care)? This was controversial. Liuba thought that it was exactly the opposite – it is easier for memory institutions, because they already have procedures. Christoph thought it was just too broad; you can always find an exception. Mary looked at the problem from the aspect of finding the data to appraise, which is the hard problem for large organizations; once the items are found, appraisal is easier. Finally, Raivo argued that the challenge is articulating the appraisal policy. This point was quite controversial! Mary and Michael totally disagreed, claiming that the actual point of difficulty is defining or implementing the procedures.

As a whole this session brought up several points of debate and pointed out that in the digital world there is a lack of clarity on the border between archiving and records management. Further, given the lack of agreement on what constitutes the hard problems, it seems like there is not a shared understanding of the problem definition.

4.2 Design for Forgetting

David Rosenthal (Stanford University Libraries, US)

Main reference P. Druschel, M. Backes, R. Tirtea, "The right to be forgotten – between expectations and practice," ENISA, [Deliverable – 2011-10-18], Nov 20, 2012.
URL https://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/the-right-to-be-forgotten

The participants in the Design for Forgetting breakout session were: Hillel Kolodner, Liuba Shrira, David Rosenthal, Lawrence You, Margo Seltzer, Mary Baker, Matthias Grawinkel, and Gerhard Schneider.

The discussion identified two reasons for forgetting data in a preservation system:

- The system may have a legal or contractual requirement to forget specific data items. For example, under the EU's "right to be forgotten", all data about an individual must be able to be removed.
- The system may need to forget data that meets specified criteria in order to meet budget or other constraints by reducing storage consumption. For example, all data more than Y years old with importance attribute less than I.

The key difference between them is the level of proof required that the data matching the criteria are gone. If the goal is to reduce resource consumption an aggregate proof, before and after resource consumption, is adequate. Different systems will need to implement different criteria and parameters for the forgetting decision algorithm; it will reduce costs if these are known at ingest time.

Google and others have found that complying with legal mandates to forget is an expensive and technically difficult process. In at least some jurisdictions it involves searching the entire content to identify the removal candidates and then performing a secure deletion process. Thus the content of the preservation system suffers frequent write updates across its entire extent. These jurisdictions apparently do not regard deleting the keys used to encrypt candidates as adequate.

The group then engaged in a lengthy discussion about the difficulty of proving that information has been destroyed. The group proposed a theorem and a corollary:

- Theorem: It is impossible to prove that information destruction has taken place. This may be a consequence of quantum determinism and reversibility, see the Black Hole Information Paradox.
- Corollary: The best achievable goal is "good enough" forgetting, such as "key tossing" – encrypting and discarding the key. Note that this makes recovering the forgotten data expensive not impossible. The cost will decrease through time.

Thus, legal interpretations surrounding "forgetting" need to be revised and clarified. They presently impose requirements that may simply be unimplementable. See for example the ENISA report.

The group was very reluctant to include forgetting as a basic requirement of preservation systems for two main reasons:

- Systems that make forgetting as hard as possible, preferably at least as hard as the printed paper library system, are important to resist censorship and protect society's heritage. See, for example, the US government's attempt to suppress Volume XXVI of Foreign Relations of the United States.
- Adding the capability to forget to a system decreases its ability to fulfill its primary mission, to preserve information. The capability's implementation may have bugs that cause unintended forgetting. Even if it is implemented perfectly, it may permit insider abuse or external attack to cause data loss that in its absence would not have occurred.

4.3 PaaS/SaaS for Data Preservation

Alexandru Iosup (TU Delft, NL)

The participants in the panel PaaS/SaaS for Data Preservation were: Jean Bacon, Christoph Becker, Nikos Chondros, Erik Elmroth, Sam Fineberg, David Giaretta, Alexandru Iosup, Natasa Milic-Frayling, Ethan Miller, Dirk Nitschke, Peter R. Pietzuch, and Raivo Ruusalepp.

How can data preservation become an application domain of cloud computing? Although cloud computing can be generically defined as a useful IT service, which type of IT service would be useful for data preservation (that is, Infrastructure-, Platform-, Software as a Service)? We propose here a set of data preservation services that together form an interface to a Platform as a Service (PaaS) cloud for data preservation. To enable both bit and logical data preservation, our proposed services store multi-layered data at the bit and logical layer. To enable long-term preservation, our services are designed to store not only raw data, but possibly also other elements in the interpretation stack, such as libraries, the OS, and even the machine model. Our approach promises to enable later re-enactment (emulation) of the recorded performance, even for complex multi-media artifacts such as online game playing and spectating. The PaaS we propose includes ten core primitive operations for data preservation, grouped into operations for ingestion, curation, and access; it also includes four primitive operations that are orthogonal to data preservation operations, such as reporting. We have validated our proposed PaaS by mapping five use cases to it: digital photography, a small company preserving financial and business records without the help of an archivist, a public museum running a digital service for the general public, a scientific setting with proprietary lab equipment and data, and digital game and game performance preservation.

5 Overview of Talks

5.1 Scientific Data Archiving Requirements

Ian F. Adams (University of California – Santa Cruz, US)

Joint work of Adams, Ian F. ; Storer, Mark W. ; Miller, Ethan L.;
Main reference I. Adams, M.W. Storer, E.L. Miller, "Analysis of Workload Behavior in Scientific and Historical Long-Term Data Repositories," ACM Trans. on Storage, Vol. 8, No. 2, ACM, 2012.
URL http://dx.doi.org/10.1145/2180905.2180907

Scientific data archiving is one of many important areas within the growing field of archival storage and has several defining characteristics to note. Its data sets may be quite large in private or controlled data, but tend to be more modest in size for public or web accessible data. In all cases however, we find that automated processes such as integrity checking, indexers, and file migration make up the majority of activity. We find that the old adage of "Write-Once, Read-Maybe" archival data should not be relied upon. Updates to data are infrequent compared to enterprise data, but hardly rare. We also found that while individual files are often not appreciably more popular than any other file or record in a corpus, users often show strong locality of access in their activities. This locality may be leveraged to improve the performance and efficiency of both local and remote storage.

From the perspective of utilizing public cloud storage, we have concerns on several fronts. First, large scientific data sets may be extremely expensive to store on a "public" cloud due to their large size. Second, some scientific data may be sensitive in nature and being stored on a shared infrastructure may be risky, particularly as cloud services often have limited, if any, liability. Third, as many providers charge on a per-access basis, we see strong disincentives for useful activities, such as web indexing and remote integrity checking.

5.2 Security technologies for cloud service provision.

Jean Bacon (University of Cambridge, GB)

Individuals and organisations wish to have security guarantees before they store their data on cloud services, short-term or long-term. Ideally, cloud service providers should state clearly what their guarantees are. But current contracts that must be accepted by cloud tenants before using cloud services explicitly avoid cloud-service providers' responsibility for potential failures. Even if this responsibility were included, it could be unenforceable because of the international scope of the jurisdiction.

Security technologies should be used as appropriate for the cloud and updated over time: authentication, access control, encryption, information flow control, data anonymisation and partitioning. I have worked with Ken Moody on role-based access control (RBAC) policy specification and enforcement; with Peter Pietzuch on Information Flow Control (IFC); on health and lifestyle monitoring in the PAL project. The focus has been on systems spanning multiple administrative domains but under a single national jurisdiction i.e., access control policy is specified nationally at a coarse grain and within individual administrative domains for fine-grain local detail. We have not worked under an assumption of international jurisdiction.

Issues from my previous work that are relevant to cloud-service provision:

- Quantification of risk: The consequences of loss or leakage of data should be quantified and used.
- Data: (1) UK cancer records have been gathered by law since 1971. Large fines are imposed if data is leaked. Cloud storage is being considered. A private cloud with known geographical locations is probably most appropriate.
- Data: (2) A large amount of personal health and lifestyle monitoring data is being gathered. How should this be summarised and stored?
- Data of types 1 and 2 tends to be append-mostly except for disambiguation and correction. Read access is needed for (1) long-term statistics about cancer and (2) analysis of personal health.
- Integrity is an aspect of data protection and must be ensured by security technology.
- Audit: Accesses to data should be logged and made available to the owner.
- Dynamic, run-time data-flow monitoring using IFC seems highly appropriate for use in the cloud.

5.3 Challenges for information longevity in the cloud

Christoph Becker (TU Wien, AT)

Searching for creative friction in the intersection of cloud technologies, delivery models, and digital longevity, we can build on the metaphor that digital preservation can be viewed both as interoperability over time and communication with the future. Keeping the bits safely stored is a necessary, but not sufficient condition for preserving information. The preservation field is increasingly moving forward from mere storage and preservation of the bits to an awareness of the value chain of information and knowledge. Preservation in this sense can be seen as the processes required to ensure that an artifact remains connected with the contemporary computing ecosystem. This is much more challenging than intuitively recognised, owing to the "black box phenomenon" of digital artifacts, which only become meaningful through a computed interpretation. While clouds introduce interesting opportunities of scale and flexibility, they also raise questions of control, transparency, and trust. I shortly discuss a few easily overlooked issues surrounding the obsolescence debate and raise a few opportunities and challenges:

1. How can we standardise and automate the underlying key processes of information preservation to leverage the flexible scalability of cloud technologies and use emerging delivery models for addressing the long tails of content artifacts, users, and access environments?
2. What are the basic factors contributing to life expectancy of digital artifacts? Can we predict life expectancy and life cycle costs for digital information in dynamic environments?
3. How can we integrate the concerns of digital longevity and information preservation over time into emerging computing paradigms? How can we establish longevity as a valid design concern in the information systems life cycle from the very beginning?

Taking the example of a specific "knowledge organisation", we see that on many conceptual levels, it does not matter much if we preserve images, documents, research data, health records, or videos – the underlying fundamental computing and communication principles, as well as many of the organisational questions, apply equally across content types and scenarios. A key challenge for both preservation and cloud computing over the next decade will be the question of systematic assessment of information artifacts, processes, systems, and organisational capabilities across scenarios and artifacts.

5.4 Reliability and Integrity of Cloud Storage

Andre Brinkmann (Universität Mainz, DE)

Cheap Cloud storage can become an interesting alternative to in-house archiving and preservation. Nevertheless, trust in the reliability, integrity, and security of Cloud storage still has to be built. Many of the design challenges for Cloud-based preservation have already been investigated in previous work on distributed storage environments such as LOCKSS or OceanStore. The storage should be assumed to be unreliable, intrusion detection should be integrated, and third-party reputation as well as long-term secrets should be avoided. Interestingly, the reliability concerns are only partly due to the quality of the provider's backend storage, but also based on the availability of the Internet connection.

Integrity is one of the key requirements of preservation, but could also conflict with the demand of Cloud providers to implement the storage backend as cheaply as possible. It is therefore important to regularly check the integrity of stored data, as data losses are otherwise silent for the data owner. Unfortunately, Cloud storage is charged based on accesses and transfer volume, therefore scanning the complete archive becomes too expensive. Efficient techniques from secure auditing could be used to uncover larger losses, small incidents are more difficult to find without reading huge parts of the archive.

Ensuring integrity and reliability clearly includes conflicting demands. Lots of copies make stuff expensive and using Cloud storage should help reduce costs. Furthermore, you should not trust a single provider, but unlike in previous work on distributed storage, the number of cloud providers is rather small. The protocols therefore have to work on this small set, requiring us to rethink the underlying assumptions of many distributed storage protocols. Furthermore, techniques such as secure auditing help detect data losses, but do not prevent them. Recovering from data losses therefore requires multiple sites to be involved, and these sites have to be coordinated, at best, based on interfaces agreed to by all Cloud providers.

5.5 Content verification

Nikos Chondros (University of Athens, GR)

The integrity of an archive needs verification to protect against both bit-rot and malicious modification. Due to the latter, even when doing this locally by storing hashes of the content and later verifying them, it necessitates going outside the digital preservation system and asking external nodes to be witnesses to the archive's integrity. For example, these witnesses might store the complete archive, a portion of the complete archive, or perhaps a versioned digest of a hash-tree that summarizes all content checksums. The use of external witness nodes results in the formation of a distributed system, adding to the complexity of the solution. If the digital preservation system is stored in the cloud, the new challenge is whether these external nodes are leased from the same cloud provider or not, questioning the single provider approach. Our current research focuses on implementing an efficient solution based on a persistent hash tree (RBB-Tree, Petros Maniatis), with the aim of minimizing nodes' storage requirements.

5.6 Thoughts on (Preventing) Logical Obsolescence

Michael Factor (IBM Research – Haifa, IL)

Logical obsolescence occurs when one is no longer able to interpret a digital object. This problem exists whether or not a cloud is part of the preservation infrastructure; however, the use of a cloud can most definitely exacerbate the problem by further separating the data owner from the preservation infrastructure. Based upon experience in the EU funded ENSURE project, there are several points we should consider. First, sometimes it is best to do nothing – what we do needs to depend upon the value of the data as well as the cost of any action. In this context, it is important to define a preservation plan based upon requirements, evaluating cost, risks and value and protecting different data in different ways, including stopping investment in some data after a period of time. Second, don't forget the metadata; it may be the most important thing. We need to leave ourselves *inexpensive* breadcrumbs when the information is stored. We should do this by building on OAIS, e.g., mapping OAIS APIs to objects, e.g., via the Cloud Data Management Interface (CDMI), storing OAIS metadata as object attributes and ensuring we have meaningful, cost effective, metadata for all entities at ingest. And finally, automation of action and verification is essential. In ENSURE, we identified the following as important approaches: 1) use a workflow engine, such as jBPM, to manage flow of all actions, 2), use virtual appliances which encapsulate the software used for the data to provide shorter term preservation and 3) use some form of computational storage, such as storlets, for transformations and quality verifications to support longer term preservation.

References

1 O. Edelstein, M. Factor, R. King, T. Risse, E. Salant and P. Taylor, "Evolving Domains, Problems and Solutions for Long Term Digital Preservation", in *Proceedings iPRES 2011 – 8th International Conference on Preservation of Digital Objects*, Singapore, 2011.

5.7 The Self-contained Information Retention Format (SIRF)

Sam Fineberg (HP Storage CT Office – Fremont, US)

Joint work of Fineberg, Sam; Cohen, Simona; Cummings, Roger
URL https://sites.google.com/site/snialtr/self-contained-informarion-retention-format-sirf

The SNIA Long Term Retention Technical Working Group (LTR TWG) was created to address the "grand technical challenges" of long term digital information retention & preservation, namely both physical ("bit") and logical preservation. A major component of the TWG's Program of Work is the creation of a logical container format, named the Self-contained Information Retention Format (SIRF), for the long-term storage of digital information.

Key aspects of a long term storage container:

- Self-describing – can be interpreted by different systems
- Self-contained – all data needed for the interpretation is in the container
- Extensible – so it can meet future needs

SIRF is a logical data format intended to be the digital equivalent of an archivist's box

- Logical container for a set of (digital) preservation objects and a catalog
- The SIRF catalog contains metadata related to the entire contents of the container as well as to the individual objects
- SIRF standardizes the information in the catalog

The LTR TWG is currently creating bindings of SIRF for Tape (Linear Tape File System) and the cloud (Cloud Data Management Interface). We see SIRF as a key part of a sustainable cloud preservation store.

5.8 Logical obsolescence

Matthias Grawinkel (Universität Mainz, DE)

The term obsolescence opens a diversity of questions on toddy's storage systems. When is data forgotten so that information cannot be accessed anymore?

Data can be deleted on demand, but how do we guarantee that no copies are left? Metadata, links or encryption keys may be lost, so that an orphaned file cannot be interpreted anymore or can only be interpreted at high overhead and cost. While a storage provider can prove that it has stored particular data, can it also prove that it does no longer stores any copies of the data?

Managing expiry dates and encryption keys solve only one part of the problem. Data can be spread to multiple data centers and to different storage media, so that deleting each byte that make up the file is impossible. Instead, encryption keys can be deleted, but legal issues remain.

5.9 Towards Logging and Preserving the Entire History of Distributed Systems

Alexandru Iosup (TU Delft, NL)

The implications of archiving large amounts of daily information for science and society are clear since at least the 1940s, when Vannevar Bush defined the concept of the personal memex as an individual's device for storing and accessing all information and communication involving that individual. Among these benefits are learning about and eradicating humankind diseases, enabling human beings more creative and thought-related time by eliminating tasks that can be automated, etc. Similarly, we posit that archiving large amounts of operational traces collected from the many distributed systems that currently underpin societies across the world would be beneficial for tuning today's systems and designing better systems in the future. What is the Distributed Systems Memex? How can such a Memex be designed and implemented? To address these and related questions, we have taken a bottom-up approach, in which we focus on the archival needs of specific application areas in which distributed systems are prominent and hope to gain sufficient understanding for the future.

In the mid 1990s, the grid computing community promised the "compute power grid," a utility computing infrastructure for scientists and engineers. Since then, a variety of grids have been built worldwide, for academic purposes, specific application domains, and general production work. The Grid Workloads Archive (GWA) [4][1] is a workload data exchange and a meeting point for the grid community. We have defined the requirements for building a workload archive and have described the approach taken to meet these requirements with the GWA. We have introduced a format for sharing grid workload information and the tools associated with this format. Using these tools, we have collected, ingested in the archive, and processed data from over fifteen well-known grid environments, covering thousands of scientists submitting millions of jobs over a period of over twenty operational years, and with working environments spanning hundreds of operational sites comprised of tens of thousands of processing machines.

Resource failures are currently common in distributed systems, likely as a side-effect of the increasing complexity and scale of these systems in real- life deployments. To facilitate the design, validation, and comparison of fault-tolerant models and algorithms, we have created the Failure Trace Archive (FTA) [2][2] as an online public repository of availability traces taken from diverse parallel and distributed systems. We have designed a new data format and a toolbox that facilitates automated analysis of trace data sets. We have collected, ingested in the archive, and processed data from over twenty-five well-known distributed systems, covering millions of users over a period of over twenty operational years, and with working environments spanning the entire world and millions of computers.

Peer-to-Peer (P2P) systems have gained a phenomenal popularity in the past few years; among them, BitTorrent alone serves daily tens of millions of people and generates an important fraction of the Internet traffic. Measurement data collected from real P2P systems are fundamental for gaining solid knowledge of the usage patterns and the characteristics of these systems and can improve the modeling, the design, and the evaluation of P2P

[1] http://gwa.ewi.tudelft.nl/
[2] http://fta.scem.uws.ed u.au/

systems. We have created the P2P Trace Archive (P2PTA) [3][3], an archive that facilitates the collection and exchange of P2P traces. Currently, the P2PTA hosts over twenty traces of various P2P applications (from file-sharing to VoIP to video-streaming), with over 60 million sessions, tens of millions of content items, and multiple years of system operation.

Spurred by the rapid development of the gaming industry and the expansion of Online Meta-Gaming Networks (OMGNs), we have designed the Game Trace Archive (GTA) [1][4] to be a virtual meeting space for the game community. We have proposed a unified format for game traces and introduced a number of tools associated with the format. With these tools, we have collected, processed, and analyzed 9 traces of both games and OMGNs. We have already collected in the GTA traces corresponding to more than 8 million real players and more than 200 million information items, spanning over 14 operational years. We also show that the GTA can be extended to include a variety of real-game trace types.

Our work in designing and building a Distributed Systems Memex has only just begun.

References

1 Yong Guo, Alexandru Iosup: The Game Trace Archive. NetGames 2012: 1-6

2 Derrick Kondo, Bahman Javadi, Alexandru Iosup, Dick H. J. Epema: The Failure Trace Archive: Enabling Comparative Analysis of Failures in Diverse Distributed Systems. CC-GRID 2010: 398-407. Best paper award.

3 Boxun Zhang, Alexandru Iosup, Johan Pouwelse, and Dick Epema. 2010. The peer-to-peer trace archive: design and comparative trace analysis. In Proceedings of the ACM CoNEXT Student Workshop (CoNEXT '10 Student Workshop). ACM, New York, NY, USA, Article 21.

4 Alexandru Iosup, Hui Li, Mathieu Jan, Shanny Anoep, Catalin Dumitrescu, Lex Wolters, Dick H. J. Epema: The Grid Workloads Archive. Future Generation Comp. Syst. 24(7): 672-686 (2008)

5.10 Endowment Models and Archival Institutions

Ross King (Austrian Institute of Technology – Wien, AT)

A growing number of archival institutions are turning towards POSE (Pay Once, Store Eternally) or Endowment models for funding their long-term digital archiving and preservation activities. The endowment model has a number of seductive advantages. First, it fits in nicely with project-oriented digitisation efforts, as the endowment costs can be included in a project budget and do not have to be added to annual running budgets. Endowment models also allow simple budget calculations based on total storage volume, which in turn support business models based on archival services. As archival institutions face pressure to become self-sustaining, such business models are in great demand.

However, there may be a number of dangerous assumptions behind simple endowment models. There is a pervading view that data centers with endowment models are like pension plans, in which incoming endowments (workers) will pay for old data (retirees). This analogy is clearly false because, unlike unfortunate pensioners, old data never dies. The reply to this

[3] http://p2pta.ewi.tudelft.nl /
[4] http://gta.st.ewi.tudelf t.nl/

is usually, "but the old data is so much smaller than the new data." This is true, but the only way in which an endowment model can handle the ever-increasing volumes of new data is by basing the cost model on a careful analysis of storage costs, such as detailed in [1]. Too many endowment models simply assume that per volume storage costs will continue to decrease (Kryder's law) forever, which simply cannot be the case. Rather, the storage capacity per unit cost, which is presently in an exponential growth phase, will eventually reach a stationary phase and level off, just as every other exponential growth scenario in nature. It is incumbent upon an endowment model to at least attempt to predict when this stationary phase will occur and at what rate storage capacity will continue to grow.

Commercial storage providers such as Google and Amazon are well-aware of these difficulties, and offer business models that are highly advantageous to themselves as a result (the decreases in service costs offered over the past five years are still much higher than the real decrease in storage costs). The question is, are libraries, archives, and other data centers equally aware? Endowment models are complex and probably more expensive than we think. The inevitable conclusion is that we can no longer afford to archive everything.

References

1 David S.H. Rosenthal, Daniel Rosenthal, Ethan L. Miller, Ian Adams, Mark W. Storer, Erez Zadok, "The Economics of Long-Term Digital Storage", *The Memory of the World in the Digital Age: Digitization and Preservation*, September 2012.

5.11 Implementing a Preservation System over a Storage Cloud

Hillel Kolodner (IBM Research – Haifa, IL)

There are many issues and trade offs to be considered in building preservation systems over storage clouds. Typically a storage cloud provides high levels of security and reliability, and in addition, advanced features to differentiate itself from competitors. Yet, paradoxically these features may be overkill for a preservation system. In particular, a preservation system cannot incur the risk of depending on a single cloud provider, who may go out of business, and also may not trust the provider to keeps its content secure. So, given that the preservation system is itself built over multiple, possibly untrustworthy clouds, it must take on the responsibility for high reliability and end-to-end security and can likely make due with clouds that provide a lower level of reliability and security. This is similar to the way that cloud systems, themselves, are typically built, where it does not pay to invest in high reliability at the low level, e.g., in hardware, since the high level cloud software needs to be able to deal with failure in any case.

On the other hand cloud storage systems are providing increasingly more sophisticated features that can simplify the task of building a preservation system. If preservation systems are built over multiple clouds, how will they leverage these features? For example, consider some of the advanced features provide by VISION Cloud, an FP7 project: support for rich metadata, computation in the storage system and support for the secure handling of data.

Rich metadata is supported as an integral part of an object by the storage system. The storage system is responsible for the integrity of the metadata as well as the data. Whereas object data cannot be updated in place, metadata can be updated. Furthermore, objects can be found/searched based on the values of their metadata fields.

The storage system provides a way to run computations in a safe and secure way. This can be more efficient as it can avoid network bandwidth. Furthermore, it can be more secure since the data does not need leave the storage system.

Support for the secure handling of data includes secure isolation of tenant/user data, encryption of data, and geographic constraints on the placement of data.

References

1 E. K. Kolodner, S. Tal, D. Kyriazis, D. Naor, M. Allalouf, L. Bonelli, P. Brand, A. Eckert, E. Elmroth, S. V. Gogouvitis, F. Harnik, D.and Hernandez, M. C. Jaeger, E. B. Lakew, J. M. Lopez, M. Lorenz, A. Messina, A. Shulman-Peleg, R. Talyansky, A. Voulodimos, and Y. Wolfsthal. A cloud environment for data-intensive storage services. In *Proceedings of the 2011 IEEE Third International Conference on Cloud Computing Technology and Science*, CLOUDCOM '11, pages 357–366, Washington, DC, USA, 2011. IEEE Computer Society.

5.12 The Life of Digital and the Cloud

Natasa Milic-Frayling (Microsoft Research UK – Cambridge, GB)

URL http://www.scape-project.eu/

Digital obsolescence is an ecosystem problem. Finding a solution requires understanding the nature of digital technologies and the structure of the Computing Technology Ecosystem (CTE) that has emerged to support the creation, application, and sustainability of digital technologies as well as the production and the reuse of digital assets.

Products of digital technologies are digital artifacts that are created through computation and can be consumed in its digital form only when the computation can be executed. A digital artifact cannot be stored – it is, at the most basic level, a sensory experience that is enabled by computation through adequate electronic hardware and lasts only while the computation lasts.

One important aspect of digital is reuse. We persist both the computer programs and the results of computation so that they can be re-instantiated again. We ensure that we have all the programs required to realize the digital artifact in the form we can consume. In many instances that involves a number of applications: the application for writing software, i.e., the original program, the application that can instantiate the results of the program, i.e., data files or document files. Finally, we need the complete stack of software to run the program as well as the hardware that enables us to sense, i.e., perceive, the digital artifact through viewing, hearing, or through our tactile sensors.

Preserving data files and program files is therefore not sufficient for preserving our digital assets. Persisted encoding of programs and data is necessary but far from sufficient. The existence of digital is about computation. Only through computation can it be experienced.

Thus, we define digital preservation as enabling digital artifacts to be instantiated in the contemporary computing environment. That may happen in different ways, but the key is always computation. We can create a virtual machine (VM), emulating the hardware and providing the required software stack. Alternatively we can port the software to the new environment and utilize the data files as they are. Finally, we can identify a contemporary application with required functionality and transform the data file format into the format that can be consumed by that application. This approach is referred to as format migration.

The cloud naturally arises as an environment in which access to digital content can be enabled over a long period of time. First, it brings together the persistent aspects of digital assets, i.e., data and program file storage, and the computation. To illustrate that point, we have created, as part of the SCAPE project, an Azure based service for format migration that is extendable in both directions – the data storage and transformation of formats. The key in ensuring access to digital artifacts is to provide an efficient software development layer that enables developers to create 'bridging' components. Bridging software can, for example, be a format translator that converts an old format to a contemporary format. It can be a virtual machine that enables hosting of a program and computing on the data files. If the results of computation need to be reused outside the virtual machine, then we need either format translators that run within VM or software to capture digital artifacts directly from the presentation layer and then convert them into a form that can be used with other applications outside the VM.

The cloud paradigm may ease the pain of digital obsolescence but only if we put in place the standards that ensure interoperability among cloud platforms and demand that they are designed with the longevity of digital in mind.

5.13 The Economics of Devices Over the Long Term

Ethan Miller (University of California – Santa Cruz, US)

Joint work of Miller, Ethan L.; Rosenthal, Daniel; Rosenthal, David S. H.; Adams, Ian F.; Zadok, Erez

The economics of long-term storage are different from those of "normal" storage: in long-lived systems, different metrics become important. The cost of long-term storage depends on complex, time-dependent interactions between metrics such as device lifetime, density growth rates, and long-term capital expenses, as well as costs of switching to media.

We are investigating the impact of several factors on the cost of providing long-term storage, both for single organizations and for cloud providers, since these costs must be paid either initially when content is first stored or over time to maintain content. For example, we find that the historical rapid growth in storage density makes it worthwhile to replace devices before they fail; as density growth rates slow, it becomes worthwhile to make devices more reliable. Similarly, keeping devices for longer reduces data migration costs and integrity verification costs, further motivating more reliable devices. Further, by moving preservation storage into a cloud environment, we can better spread capital costs over time and provide better device and software diversity.

5.14 Domain Specific Needs: Archives are becoming mission critical

Dirk Nitschke (Oracle – Herndon, US)

The use of archives is changing. A common perception is that (classical) archives can be *slow* and it's not a problem when they are *closed.* However, digital information changes the game in several ways:

More and more companies provide 24 × 7 services and create centralized archives instead of multiple isolated solutions. Consequently, multiple departments or applications depend on the centralized archive. This implies that a digital archive must always be up and running.

Companies store their most valuable assets in these archives. The digital assets are created by a variety of different applications, and very often they are the source material for other products. Data loss is not an option, and the archive must integrate into your enterprise applications.

The end user expectation in regards of access latency changes. Today, we are all used to getting answers to our questions instantaneously from our favorite Internet search engine, and we can download data at a reasonable bandwidth. The same is expected from archives today.

Data is ingested into and retrieved from company archives by end users, not by specialized personal, so they must be easy to use.

Nobody claims that this is easy to accomplish, nor is there a one size fits all solution.

If you are involved in an archiving project, make sure to talk to the right people. Archiving is an organizational task and not an IT task. Typically, the IT department has no knowledge about the data that they have to store. Define who wants to archive what, why, and how you intend to re-use the data in the future. Afterwards, do a data classification. Find out what you have, evaluate the value of your data over time and make some decisions. Making decisions can be the hardest part because you have to decide what *not* to archive.

Then you should think about your time scale, how long to keep the data, the projected amount of data and objects and your migration strategy.

Last but not least, select the right tools and do not over-engineer. Someone has to run the system for a long period of time.

5.15 Cloud computing and future memory

Gillian Oliver (Victoria University – Wellington, NZ)

The objectives of cultural heritage institutions are concerned with individual, organisational and societal memory, but it is important to note that memory consists of remembering and forgetting, plus has long term and short term dimensions. Accordingly, Information professionals in workplaces and memory institutions have developed tools and techniques to prioritise and target preservation actions (including destruction). The advent of cloud computing, and the consequent outsourcing of memory, poses significant challenges to our future memory. Future memory could be a massive accumulation of digital information sludge, with remnants of what should be forgotten remaining, and only occasional glimpses of the important and truly significant. Consequently archival authorities around the world have identified risks associated with cloud computing and are actively working to raise awareness of the issues involved.

5.16 Long-Term Cloud Storage needs Data-Centric Security

Peter R. Pietzuch (Imperial College London, GB)

Joint work of Pietzuch, Peter R.; Bacon, Jean; Moody, Ken

Security considerations are a major issue holding back the widespread adoption of cloud storage: many organisations are concerned about the confidentiality and integrity of their users' data when hosted in third-party public clouds. Today's cloud storage providers struggle to give strong security guarantees that user data belonging to cloud tenants will be protected "end-to-end", i.e. across its entire life cycle. Therefore security engineering must be integrated with all stages of data storage in clouds. We want cloud providers to isolate data of each of their clients. This is crucial for cloud infrastructures, in which the stored data have different owners whose interests are not aligned (and may even be in competition).

We propose a principled approach to designing and deploying *end-to-end secure data storage* in the cloud by means of thorough tagging of the security meaning of data, analogous to what is already done for data types [1]. The aim is that such a *data-centric* security approach using Information Flow Control (IFC) techniques can ensure that—above a small trusted code base—data cannot be leaked by buggy or malicious software. End-to-end information flow control thus preempts worries about security and privacy violations. The cloud storage infrastructure enforces data flow policies through multiple layers of security mechanisms following a *defense-in-depth* strategy: based on policies, it creates *data compartments* that isolate user data. A small privileged kernel, which is part of the cloud infrastructure, constitutes a trusted computing base (TCB), and tracks the flow of data between compartments, preventing data flows that would violate policies. Due to its minimal size and reliance on hardware protection mechanisms, such an approach can strengthen the security of a cloud storage infrastructure against internal and external security attacks.

References

1 Jean Bacon, David Evans, David M. Eyers, Matteo Migliavacca, Peter Pietzuch, and Brian Shand, "Enforcing End-to-end Application Security in the Cloud", ACM/IFIP/USENIX 11th International Middleware Conference (Middleware'10), Bangalore, India, November 2010.

5.17 Using Provenance to Protect Data

Margo Seltzer (Harvard University, US)

When we discuss protecting the data, we typically make the assumption that we are protecting the data from adversarial attack and that the only thing of import is the "integrity" of the data, for some definition of integrity. However, I believe that the question is broader. There are (at least) three different constituencies to whom we can offer protection: the owner of the data, the user of the data, and the provider of the data. Provenance or lineage, which is the complete history of how the data came to be in its current form and at its particular location, is critical for all parties.

For the data provider, data provenance documents the authenticity of the data and ensures that the provider has the appropriate rights to store and serve the data. For the user, authenticity is a key concern, but so too are details of the processing and transformation that have been applied to the data. A user may want to know specific versions of software used to analyze data or the particular system on which a data set was produced. Finally, a data owner uses provenance to establish or maintain reputation as well as to track the data as it is disseminated. All three constituencies have a vested interest in maintaining complete and reliable provenance.

Maintaining reliable provenance requires cooperation between all the systems that participate in data creation, transmission, and storage. While such cooperation would seem to imply that we are paralyzed without well-established standards, I claim that we cannot let ourselves be paralyzed. Data is being produced, manipulated, transmitted, stored, copied, transformed, and destroyed constantly. We need to start documenting that now – we should accept the fact that different data will have different provenance and establish simple ways to communicate the provenance from users to systems, systems to systems, and systems to users.

From an archive's point of view, it must be able to accept provenance from an external source, add provenance to record all archival acts, integrate seamlessly with external and internal sources, and not require that all systems agree on a single, standard format/representation.

5.18 Modularity and incrementalilty in handling logical obsolescence

Liuba Shrira (Brandeis Univ. Waltham, US)

The software stack in a cloud-based preservation service will have different layers. Each layer will need to handle the logical obsolescence problem. We hypothesize that similar properties will be desirable in different layers, ideally allowing us to use similar techniques to achieve them.

With this goal in mind, this talk puts forward an abstract framework for handling logical obsolescence for long-lived interlinked stateful objects described by a schema, borrowed from our work on a system for automatic data store object upgrades. In such a system objects may need to be migrated in two directions, from older version to newer version to allow us to ask new questions about old data, and from newer version to old version to allow us to ask old questions about new data.

We consider two desirable upgrade properties, modularity, which makes it easy to write code that automatically transforms objects from one version to another, and incrementality, which enables low-cost on-demand object migration between versions, and discuss the difficulties of achieving them, hoping to examine with the group how these concerns apply to the obsolescence handling across the different layers.

Participants

- Ian F. Adams
  University of California – Santa Cruz, US
- Jean Bacon
  University of Cambridge, GB
- Mary Baker
  HP Labs – Palo Alto, US
- Christoph Becker
  TU Wien, AT
- André Brinkmann
  Universität Mainz, DE
- Nikos Chondros
  University of Athens, GR
- Milena Dobreva
  University of Malta, MT
- Erik Elmroth
  University of Umeå, SE
- Michael Factor
  IBM – Haifa, IL
- Sam Fineberg
  HP Storage CT Office – Fremont, US
- David Giaretta
  APA, Dorset, GB
- Matthias Grawinkel
  Universität Mainz, DE
- Alexandru Iosup
  TU Delft, NL
- Ross King
  Austrian Institute of Technology – Wien, AT
- Hillel Kolodner
  IBM – Haifa, IL
- Ewnetu Bayuh Lakew
  University of Umeå, SE
- Natasa Milic-Frayling
  Microsoft Research UK – Cambridge, GB
- Ethan Miller
  University of California – Santa Cruz, US
- Dirk Nitschke
  Oracle – Herndon, US
- Gillian Oliver
  Victoria Univ. – Wellington, NZ
- Peter R. Pietzuch
  Imperial College London, GB
- David S. H. Rosenthal
  Stanford University Libraries, US
- Raivo Ruusalepp
  National Library of Estonia – Tallinn, EE
- Gerhard Schneider
  Universität Freiburg, DE
- Margo Seltzer
  Harvard University, US
- Liuba Shrira
  Brandeis Univ. Waltham, US
- Joanne Syben
  Google Inc. – Mountain View, US
- Lawrence You
  Google Inc. – Mountain View, US

Report from Dagstuhl Seminar 12481

Quantitative Security Analysis

Edited by

Boris Köpf[1], Pasquale Malacaria[2], and Catuscia Palamidessi[3]

1 **IMDEA Software Institute, ES**
2 **Queen Mary University of London, GB,** pm@dcs.qmw.ac.uk
3 **Ecole Polytechnique – Palaiseau, FR,** catuscia@lix.polytechnique.fr

Abstract

The high amount of trust put into today's software systems calls for a rigorous analysis of their security. Unfortunately, security is often in conflict with requirements on the functionality or the performance of a system, making perfect security an impossible or overly expensive goal. Under such constraints, the relevant question is not whether a system is secure, but rather how much security it provides. Quantitative notions of security can express degrees of protection and thus enable reasoning about the trade-off between security and conflicting requirements. Corresponding quantitative security analyses bear the potential of becoming an important tool for the rigorous development of practical systems, and a formal foundation for the management of security risks.

Seminar 20.–25. November, 2012 – www.dagstuhl.de/12481
1998 ACM Subject Classification C.2.0 General: Security and protection, E.4 Coding and Information Theory, D.3 Programming Languages, D.2.4 Software/Program Verification: Formal methods
Keywords and phrases Security, Privacy, Information theory, Programming languages, Formal methods
Digital Object Identifier 10.4230/DagRep.2.11.135

1 Executive Summary

Boris Köpf
Pasquale Malacaria
Catuscia Palamidessi

The high amount of trust put into today's software systems calls for a rigorous analysis of their security. Unfortunately, security is often in conflict with requirements on the functionality or the performance of a system, making perfect security an impossible or overly expensive goal. Under such constraints, the relevant question is not whether a system is secure, but rather how much security it provides. Quantitative notions of security can express degrees of protection and thus enable reasoning about the trade-off between security and conflicting requirements. Corresponding quantitative security analyses bear the potential of becoming an important tool for the rigorous development of practical systems, and a formal foundation for the management of security risks.

While there has been significant progress in research on quantitative notions of security and tools for their analysis and enforcement, existing solutions are still partial. The focus of the seminar is to discuss the following key issues.

Quantitative Security Analysis, *Dagstuhl Reports*, Vol. 2, Issue 11, pp. 135–154
Editors: Boris Köpf, Pasquale Malacaria, and Catuscia Palamidessi
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Quantitative Notions of Security: A single qualitative security property may give rise to a spectrum quantitative generalizations, each with different characteristics and application domains. For quantitative confidentiality, current research focuses on differential privacy and measures based on information-theoretic entropy. For other security properties such as integrity, availability, incoercibility, vote verifiability, etc., quantitative generalizations are only now emerging or have not even been proposed. One goal of this seminar is to advance the understanding of the relationship between existing quantitative security properties, and to join forces in the development of new ones.

Tools for Quantitative Security Analysis: Performing a quantitative security analysis of a realistic system is a challenging problem due to the complexity of modern software. It is mandatory to provide developers with tool support for this task. One goal of this seminar is to advance the understanding of the fundamental reasoning principles for quantitative notions of security, their connection to programming languages and verification techniques, and the theoretical limits for automatically deriving quantitative security guarantees.

Novel Application Domains: Quantitative security analyses have been successfully applied, e.g., for quantifying the side-channel leakage in cryptographic algorithms, for capturing the loss of privacy in statistical data analysis, and for quantifying security in anonymity networks. In emerging application domains such as electronic voting or distributed usage control, the need for quantitative analyses has been recognized. It is a goal of this seminar to foster the collaboration between experts in emerging application domains and those in quantitative security analysis.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 Not all bits are created equal: incorporating the meaning and value of secret bits into measures of information flow

Mario Alvim (University of Pennsylvania, US)

Joint work of Alvim, Mário S.; Scedrov, Andre; Schneider, Fred B.

Most established information-theoretic approaches to quantitative information flow (QIF) define information in terms of Shannon entropy, min-entropy, or guessing entropy, which are a measure of, respectively, how much information flows, how likely it is that the secret be guessed in one try, and how long it takes to the secret to be guessed. These measures implicitly assume that every bit of the secret has the same "value", and therefore that every leaked bit represents the same 'threat". In many practical scenarios, however, some bits represent more important information than others (e.g. in a bank system, the bits representing the client's account number and pin code are more sensitive - and valuable – than the bits representing the client's street address). In this talk we discuss ongoing work on how to incorporate the "value" (i.e. meaning) of bits into measures of information and leakage. We consider deterministic systems, and use the Lattice of Information as an underlying algebraic structure for the set of all possible attacks an adaptive adversary can perform. We propose several measures for the information carried by the elements in the lattice. In particular, we are able to show that the measures for QIF based on Shannon entropy, min-entropy and guessing entropy are a special case of our approach, where every field of the secret is considered to be equally valuable.

3.2 Multi-run security

Arnar Birgisson (Chalmers UT – Göteborg, SE)

Joint work of Birgisson, Arnar; Sabelfeld, Andrei
Main reference A. Birgisson, A. Sabelfeld, "Multi-run security," in Proc. of the European Symp. on Research in Computer Security (ESORICS'11), LNCS, Vol. 6879, pp. 372–391, Springer, 2011.
URL http://dx.doi.org/10.1007/978-3-642-23822-2_21
URL http://www.hvergi.net/arnar/publications/pdf/esorics11.pdf

This paper explores information-flow control for batch-job programs that are allowed to be re-run with new input provided by the attacker. We argue that directly adapting two major security definitions for batch-job programs, termination-sensitive and termination-insensitive noninterference, to multi-run execution would result in extremes. While the former readily scales up to multiple runs, its enforcement is typically over-restrictive. The latter suffers from insecurity: secrets can be leaked in their entirety by multiple runs of programs that are secure according to batch-job termination-insensitive noninterference.

Seeking to avoid the extremes, we present a framework for specifying and enforcing multi-run security in an imperative language. The policy framework is based on tracking the attacker's knowledge about secrets obtained by multiple program runs. Inspired by previous work on robustness, the key ingredient of our type-based enforcement for multi-run security is preventing the dangerous combination of attacker-controlled data and secret data from affecting program termination.

3.3 Worst- and average-case privacy breaches in randomization mechanisms

Michele Boreale (Università di Firenze, IT)

Joint work of Boreale, Michele; Paolini, Michela
Main reference M. Boreale, M. Paolini, "Worst- and average-case privacy breaches in randomization mechanisms," in Proc. of IFIP-TCS 2012, LNCS, Vol. 7604, pp. 72–86, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-33475-7_6

In a variety of contexts, randomization is regarded as an effective technique to conceal sensitive information. We model randomization mechanisms as information-theoretic channels. Our starting point is a semantic notion of security that expresses absence of any privacy breach above a given level of seriousness ϵ, irrespective of any background information, represented as a prior probability on the secret inputs. We first examine this notion according to two dimensions: worst vs. average case, single vs. repeated observations. In each case, we characterize the security level achievable by a mechanism in a simple fashion that only depends on the channel matrix, and specifically on certain measures of "distance" between its rows, like norm-1 distance and Chernoff Information. We next clarify the relation between our worst-case security notion and differential privacy (dp): we show that, while the former is in general stronger, the two coincide if one confines to background information that can be factorised into the product of independent priors over individuals. We finally turn our attention to expected utility, in the sense of Ghosh et al., in the case of repeated independent observations. We characterize the exponential growth rate of any reasonable utility function. In the particular case the mechanism provides ϵ-dp, we study the relation of the utility rate with ϵ: we offer either exact expressions or upper-bounds for utility rate that apply to practically interesting cases, such as the (truncated) geometric mechanism.

3.4 Measuring Information Leakage using Generalized Gain Functions

Kostas Chatzikokolakis (Ecole Polytechnique – Palaiseau, FR)

Joint work of Alvim, Mario; Chatzikokolakis, Kostas; Palamidessi, Catuscia; Smith, Geoffrey
Main reference M.S. Alvim, K. Chatzikokolakis, C. Palamidessi, G. Smith, "Measuring Information Leakage using Generalized Gain Functions," in Proc. of the 25th IEEE Computer Security Foundations Symposium (CSF'12), pp. 265–279, IEEE CS, 2012.
URL http://dx.doi.org/10.1109/CSF.2012.26
URL http://www.lix.polytechnique.fr/~kostas/papers/csf12_gleakage.pdf

This talk introduces g-leakage, a rich generalization of the min-entropy model of quantitative information flow. In g-leakage, the benefit that an adversary derives from a certain guess about a secret is specified using a gain function g. Gain functions allow a wide variety of operational scenarios to be modeled, including those where the adversary benefits from guessing a value close to the secret, guessing a part of the secret, guessing a property of the secret, or guessing the secret within some number of tries. I will discuss important properties of g-leakage, including bounds between min-capacity, g-capacity, and Shannon capacity. Moreover I will discuss a connection between a strong leakage ordering on two channels, C_1 and C_2, and the possibility of factoring C_1 into C_2C_3, for some C_3. Based on this connection, I will propose a generalization of the Lattice of Information from deterministic to probabilistic channels.

3.5 Estimating Information Leakage from Trial Runs and Whole Java Programs.

Tom Chothia (University of Birmingham, GB)

In this talk I will outline some results that make it possible to estimate measures of information leakage based on mutual information and min-entropy from trial runs of a system alone. We propose statistical estimation as a method of applying more theoretical work on quantitative security directly to implemented systems, and we will demonstrate this by measuring the information leak in MIX nodes and from encrypted Tor traffic. We will then present a model of leakage for complete, probabilistic, non-terminating programs and show how we can use this to estimate the information leakage from large Java Programs.

3.6 Approximation and Relative Entropy

Alessandra Di Pierro (Università degli Studi di Verona, IT)

Joint work of Di Pierro, Alessandra; Hankin, Chris; Wiklicky, Herbert

Program analysis produces approximated results due to the abstraction on the state space which is required to construct 'simplified' (computable) semantics. In the case of probabilistic abstraction, i.e. when the abstract domain is a probability space, this approximation can be seen as the 'inefficiency' of mistakenly assuming that the behaviour of a source program P is a distribution x when the true distribution is y. In terms of information theory this is represented by the notion of relative entropy, aka Kullback-Leibler divergence [5]. Based on this intuition, we re-visit the notion of Approximate Confinement introduced in [4, 3, 2]. This notion formalises probabilistic non-interference in terms of process indistinguishability according to some abstract semantics (I/O observables, bisimulation etc.) and allows for the leakage of a certain amount epsilon of information. Such a quantity corresponds to a measure of the approximation introduced by the abstract semantics (probabilistic observables) and can thus be interpreted as a measure of the KL-divergence of the system. The statistical interpretation [1] of epsilon as an estimate of the number of tests needed to differentiate two executions of a program on sensitive data (i.e. how hard an attacker has to work in order to breach security) is also in accordance with the hypothesis testing formulation of relative entropy.

References

1 J. Shao, Mathematical Statistics, Springer Texts in Statistics, Springer Verlag, New York – Berlin – Heidelberg, 1999.

2 A. Di Pierro, C. Hankin, H. Wiklicky, Measuring the confinement of probabilistic systems, Theoretical Computer Science 340 (1) (June 2005) 3–56

3 A. Di Pierro, C. Hankin, H. Wiklicky, Approximate Non-Interference, Journal of Computer Security 12 (1) (2004) 37–81.

4 A. Di Pierro, C. Hankin, H. Wiklicky, Approximate non-interference, in: Proceedings of CSFW'02, IEEE, Cape Breton, Canada, 2002, pp. 3–17.

5 S. Kullback, R. A. Leibler, On Information and Sufficiency, Ann. Math. Statist. 22 (1) (1951) 79-86.

3.7 A differentially private mechanism of optimal utility for a region of priors

Ehab ElSalamouny (Ecole Polytechnique – Palaiseau, FR)

Joint work of ElSalamouny, Ehab; Chatzikokolakis, Konstantinos; Palamidessi, Catuscia
Main reference E. ElSalamouny, K. Chatzikokolakis, C. Palamidessi, "A differentially private mechanism of optimal utility for a region of priors," in Proc. of the 2nd Int'l Conf. on Principles of Security and Trust (POST'13), LNCS, Vol. 7796, pp. 41–62, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-36830-1_3
URL http://hal.inria.fr/docs/00/77/20/99/PDF/main.pdf

The notion of differential privacy has emerged in the area of statistical databases as a measure of protection of the participants sensitive information, which can be compromised by selected queries. In this talk I consider mechanisms which satisfy differential privacy by perturbing query outputs, and therefore reduce their utility. Since for any non-counting query there is no such a mechanism that is optimal for every prior (side knowledge), I highlight for an arbitrary query and a privacy parameter a special region of priors for which an optimal mechanism may exist. For each prior in this region, I show upper bounds for utility as well as for min-mutual information between the real query results and the noisy outputs reported to the user. Then, I describe a special mechanism, called the "tight- constraints mechanism", and discuss the conditions for its existence. This mechanism has the property of reaching the bounds for all the priors of the region, and thus it is optimal on the whole region. Finally I show that the same analysis implies tight upper-bounds for the min-entropy leakage about the database through any differentially private mechanism.

3.8 Quantifying Leakage in the Presence of Unreliable Source of Information

Sardaouna Hamadou (Ecole Polytechnique – Palaiseau, FR)

The frequent inaccurate, misleading or outdated material about people and businesses on social networks, online forums, blogs and other forms of online communication and information sharing raises important reputation and privacy issues.

In this talk, we will address these issues by providing a formal framework generalizing current methods of quantifying information flow. We will refine these models by integrating the notion of belief. The idea is that the threat should be relative to the possible initial belief, that is a (potentially inaccurate) information, that an attacker may have about the confidential information. More precisely, we will consider the case where the adversary is combining information from an external and potentially unreliable source and the observables of a program/protocol in order to increase her chance of breaking the privacy. In such context, concepts from Belief Theory have proved quite useful as it has higher ability to combine information from (partially or totally) disagreeing sources.

3.9 Attack Time Analysis

Holger Hermanns (Universität des Saarlandes, DE)

Joint work of Hermanns, Holger; Pulungan, Reza; Stoelinga, Marielle; Arnold, Florian

Security attacks are a threat to an increasing number of systems on which our society depends. Coined by Bruce Schneier, attack trees are a convenient graphical formalism to structure the understanding of potential security attacks and to quantify security risks.

An important concern in quantitative risk analysis is the quality of the data: how realistic are the probabilities attached to basic attack steps? In this work, we show how fitting techniques for phase type distributions can help in a time dependent risk analysis. The approach we propose combines this information with the attack tree, and turns it into a Markov chain. Compositional compression techniques are used to keep the size of this model manageable. The quantitative evaluation of this model is delegated to a stochastic model checker.

We apply this approach to a genuine attack tree example. This example reveals obvious further operators that seem natural to be included in an attack tree formalism.

3.10 Dynamic enforcement of knowledge-based security policies using probabilistic abstract interpretation

Michael Hicks (University of Maryland – College Park, US)

Joint work of Hicks, Michael; Mardziel, Piotr; Magill, Stephen; Srivatsa, Mudhakar
Main reference P. Mardziel, S. Magill, M. Hicks, M. Srivatsa, "Dynamic Enforcement of Knowledge-based Security Policies using Probabilistic Abstract Interpretation," (extended version of CSF'11 paper), October, 2012.
URL http://www.cs.umd.edu/~mwh/papers/mardziel12belieflong.html

This work explores the idea of knowledge-based security policies, which are used to decide whether to answer queries over secret data based on an estimation of the querier's (possibly increased) knowledge given the results. Limiting knowledge is the goal of existing information release policies that employ mechanisms such as noising, anonymization, and redaction. Knowledge-based policies are more general: they increase flexibility by not fixing the means to restrict information flow.

We enforce a knowledge-based policy by explicitly tracking a model of a querier's belief about secret data, represented as a probability distribution, and denying any query that could increase knowledge above a given threshold. We implement query analysis and belief tracking via abstract interpretation, which allows us to trade off precision and performance through the use of abstraction, while maintaining soundness. We have developed an approach to augment standard abstract domains to include probabilities, and thus define distributions. We focus on developing probabilistic polyhedra in particular, to support numeric programs. While probabilistic abstract interpretation has been considered before, our domain is the first whose design supports sound conditioning, which is required to ensure that estimates of a querier's knowledge are accurate.

Experiments with our implementation show that several useful queries can be handled efficiently, particularly compared to exact (i.e., sound) inference involving sampling. We also show that, for our benchmarks, restricting constraints to octagons or intervals, rather than full polyhedra, can dramatically improve performance while incurring little to no loss in precision.

Finally, I will sketch a generalization of our ideas to reasoning about information release in secure multiparty computations.

3.11 A Framework for Extracting Semantic Guarantees from Privacy Definitions

Daniel Kifer (Penn State University – University Park, US)

Joint work of Lin, Bing-Rong; Kifer, Daniel
Main reference B.-R. Lin, D. Kifer, "A Framework for Extracting Semantic Guarantees from Privacy," arXiv:1208.5443v1 [cs.DB].
URL http://arxiv.org/abs/1208.5443

The goal of statistical privacy is to choose an algorithm whose input is a sensitive data set and whose output contains useful and nonsensitive statistical information.

Privacy definitions specify the algorithms that can be used. However there is a long history of finding new weaknesses in existing privacy definitions. Thus there is a need for tools for analyzing privacy definitions. One common approach is to invent an attack and see if certain pieces of information can be inferred. This is hit-or-miss: if an attack does not work then no conclusions can be drawn, and it will not identify cases where the privacy definition protects unnecessary pieces of information.

In this talk we present a new framework for analyzing privacy definitions and deriving their semantic guarantees.

3.12 How We Should Measure the Utility of Sanitizing Mechanisms, How We Should Process Sanitized Data, and Why.

Daniel Kifer (Penn State University – University Park, US)

Joint work of Lin, Bing-Rong; Kifer, Daniel

Utility measures are used in statistical privacy to identify algorithms that produce the most useful output subject to privacy constraints. In this talk we analyze utility measures that are commonly used in the literature and show that they reward algorithms for throwing away information.

To fix this problem, we consider 3 axioms that utility measures should satisfy. Surprisingly, these axioms imply that utility should be measured using Bayesian Decision Theory even though there is nothing Bayesian about those axioms (in fact, they are quite different from the usual derivations of Bayesian decision theory).

These results imply that choosing sanitizing algorithms to maximize utility is the same as choosing sanitizing algorithms whose outputs are best analyzed using decision theoretical tools (as opposed to other statistical procedures). We conduct an experimental evaluation on the privacy-preserving sorted histogram problem and empirically show that the decision-theoretic tools consistently produce more accurate estimators than previous approaches.

3.13 Automatic Quantification of Cache Side Channels

Boris Köpf (IMDEA Software Institute, ES)

Joint work of Doychev, Goran; Feld, Dominik; Köpf, Boris; Mauborgne, Laurent; Ochoa, Martin; Reineke, Jan
Main reference B. Köpf, L. Mauborgne, M. Ochoa, "Automatic Quantification of Cache Side Channels," in Proc. of the 24th Int'l Conf. on Computer-Aided Verification (CAV'12), LNCS, Vol. 7358, pp. 564–580, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-31424-7_40

This talk presents work on a novel approach for automatically deriving upper bounds on the amount of information about the input that an adversary can extract from a program by observing the CPU's cache behavior.

Technically, our approach builds on the observation that (an upper bound on) the number of possible side-channel observations corresponds to (an upper bound on) the number of leaked bits. Such upper bounds can be obtained by computing super-sets of the set of possible observations by abstract interpretation, and by determining sizes [3]. We apply this idea to the problem of quantifying cache side-channels by providing abstract domains that keep track of the side-channel observations different kinds of cache adversaries can make, together with counting procedures for the number of corresponding concretisations.

The first part of this talk presents an initial case study, in which we combine existing tools for static cache analysis [2, 1] with a novel counting procedure for cache states. We use this combination for deriving bounds on the leakage of executables of standard AES implementations, demonstrating that automatically deriving security guarantees against cache attacks is indeed feasible. However, the obtained bounds hold only for a certain class of adversaries (namely: access-based), and their derivation requires code instrumentation.

The second part of this talk presents ongoing work on a dedicated tool for the automatic quantification of cache side-channels. The tool is based on an abstract interpretation engine for x86 binaries and can be easily extended by abstract domains for different kinds of cache observations. We provide a novel set of such abstract domains that cover all kinds of adversary models that are typically considered in the literature, namely: access-based, trace-based, and time-based. The talk concludes with experimental results, including the first security proof of the preloading countermeasure, based on an actual x86 executable of AES.

Talk is based on joint work with Goran Doychev, Dominik Feld, Laurent Mauborgne, Martin Ochoa, and Jan Reineke.

References

1 AbsInt aiT Worst-Case Execution Time Analyzers. http://www.absint.com/a3/

2 C. Ferdinand, F. Martin, R. Wilhelm, and M. Alt. Cache behavior prediction by abstract interpretation. *Science of Computer Programming*, 35(2):163 – 189, 1999.

3 B. Köpf and A. Rybalchenko. Approximation and Randomization for Quantitative Information-Flow Analysis. In *CSF*, pages 3–14. IEEE, 2010.

3.14 Algebraic Foundations for Quantitative Information Flow

Pasquale Malacaria (Queen Mary University of London, GB)

Several mathematical ideas have been investigated for Quantitative Information Flow. Information theory, probability, guessability are the main ideas in most proposals. They aim to quantify *how much information* is leaked, *how likely is to guess* the secret and *how long does it take* to guess the secret respectively. In this work we investigate the relationship between these ideas in the context of the quantitative analysis of deterministic systems. We propose the Lattice of Information as a valuable foundation for these approaches; not only it provides an elegant algebraic framework for the ideas, but also to investigate their relationship. In particular we will use this lattice to prove some results establishing order relation correspondences between the different quantitative approaches. The implications of these results w.r.t. recent work in the community is also investigated.

While this work concentrates on the foundational importance of the Lattice of Information its practical relevance has been recently proven, notably with the quantitative analysis of Linux kernel vulnerabilities. Overall we believe these works set the case for establishing the Lattice of Information as one of the main reference structure for Quantitative Information Flow.

The talk is based on the forthcoming paper [1].

References

1 P. Malacaria. Algebraic Foundations for Quantitative Information Flow. Mathematical Structures in Computer Science (To appear)

3.15 Denotational models for non-interference, probability and nondeterminism

C. Carroll Morgan (UNSW – Sydney, AU)

Joint work of McIver, Annabelle K.; Meinicke, Larissa A.; Morgan, C. Carroll
Main reference A. McIver, L. Meinicke, C. Morgan, "Compositional closure for Bayes Risk in probabilistic noninterference," in Proc. of the 37th Int'l Colloquium on Automata, Languages and Programming – Part II (ICALP'10), LNCS, Vol. 6199, pp. 223–235, Springer, 2010.
URL http://dx.doi.org/10.1007/978-3-642-14162-1_19

Combining the three features of the title is a notoriously hard problem that has attracted much interesting work. Actually the fourth feature is "denotational", as producing a denotational semantics (rather than only a mathematical model) introduces novel and challenging further constriants.

Markov Processes (for probability without nondeterminism or hiding), Markov Decision Processes (add nondeterminism) and Partially Observable MDP's (add hiding) are mathematical models. Adding the further constraints of full abstraction and compositionality, encouraged by a computer-science perspective, suggest quotients on these models that induce interesting semantic domains built from monads and that include novel "security refinement" partial orders.

I will describe the above motivations, briefly, and then in more detail describe key features of the models we have constructed with all those constraints in mind. The target for the talk will be to explain the steps we are taking with our most recent work, and to elicit suggestions from the group about possible ways forward.

As part of our contribution we proved the outstanding "Coriaceous Conjecture", an open problem due to Alvim, Chatzikokolakis, Palamidessi and Smith (also participants at the seminar). See the Seminar-Wide materials section.

The work is joint, with Annabelle McIver (Macquarie University) and Larissa Meinicke (University of Queensland).

References

1 McIver, AK, Meinicke, LA, Morgan CC. *Compositional closure for Bayes Risk in probabilistic noninterference*. Proc. ICALP 2010. http://dx.doi.org/10.1007/978-3-642-14162-1_19

2 Morgan, CC. *Compositional noninterference from first principles*. Formal Aspects of Computing 24(1):1-24. Springer (2010) http://dx.doi.org/10.1007/s00165-010-0167-y

3 McIver, AK, Meinicke, LA, Morgan CC. *A Kantorovich-Monadic Powerdomain for Information Hiding, with Probability and Nondeterminism*. Proc. LiCS 2012. IEEE (2012). http://dx.doi.org/10.1109/LICS.2012.56

3.16 Probabilistic model checking and PRISM

Gethin Norman (University of Glasgow, GB)

Main reference M. Kwiatkowska, G. Norman, D. Parker, "PRISM 4.0: Verification of Probabilistic Real-time Systems," in Proc. of 23rd Int'l Conf. on Computer Aided Verification (CAV'11), LNCs, Vol. 6806, pp. 585–591, Springer, 2011.
URL http://dx.doi.org/10.1007/978-3-642-22110-1_47

Probabilistic model checking has established itself as a valuable technique for the formal modelling and analysis of systems that exhibit stochastic behaviour. It has been applied to of a wide range of systems, from communication and security protocols to biological signalling pathways. This talk gives an overview of my research in this area, reviewing the different the models, properties and application domains that have been investigated. In addition, I will outline some techniques being developed to allow for the efficient analysis of complex and infinite state systems, as well as some open problems and future challenges.

3.17 Indistinguishable regions in Geographic Privacy

Martin Ochoa (Siemens – München, DE)

Joint work of Cuellar, Jorge; Ochoa, Martin; Rios, Ruben
Main reference J. Cuellar, M. Ochoa, R. Rios, "Indistinguishable regions in Geographic Privacy," in Proc. of the 27th Annual ACM Symp. on Applied Computing (SAC'12), pp. 1463–1469, ACM, 2012.
URL http://dx.doi.org/10.1145/2245276.2232010

The ubiquity of positioning devices poses a natural security challenge: users want to take advantage of location-related services as well as social sharing of their position but at the same time have security concerns about how much information should be shared about their exact position. This talk discusses different location-privacy problems, their formalization and the novel notion of indistinguishability regions that allows one to proof that a given obfuscation function provides a good trade-off between location sharing and privacy.

3.18 Enhancing Differential Privacy: from Hamming to General Metrics

Catuscia Palamidessi (Ecole Polytechnique – Palaiseau, FR)

Differential Privacy is one of the most prominent frameworks used to deal with disclosure prevention in statistical databases. Differential privacy is a formal privacy guarantee that ensures that sensitive information relative to individuals cannot be easily inferred by disclosing answers to aggregate queries. If two databases are adjacent, i.e. differ only for an individual, then querying them should not allow to tell them apart by more than a certain factor. The transitive application of this property induces a bound also on the distinguishability of two generic databases, which is determined by their distance on the Hamming graph of the adjacency relation.

In this paper we lift the restriction relative to the Hamming graphs and we explore the implications of differential privacy when the indistinguishability requirement depends on an arbitrary notion of distance. We show that we can express, in this way, (protection against) kinds of privacy threats that cannot be naturally represented with the standard notion. We give an intuitive characterization of these threats in terms of Bayesian adversaries, which generalizes the characterization of (standard) differential privacy from the literature. Next, we revisit the well-known result on the non-existence of universally optimal mechanisms for any query other than counting queries. We show that in our setting, for certain kinds of distances, there are many more queries for which universally optimal mechanisms exist: Notably sum, average, and percentile queries. Finally, we show some applications in various domains: statistical databases where the units of protection are groups (rather than individuals), geolocation, and smart metering.

3.19 Towards an SMT-based approach for Quantitative Information Flow

Quoc Sang Phan (Queen Mary University of London, GB)

Joint work of Phan, Quoc Sang; Malacaria, Pasquale

Quantitative Information Flow (QIF) is a powerful approach to analyse leaks of confidential information in a software system. Here we present a novel method for automated QIF analysis. We cast the problem of QIF analysis into the problem of counting boolean abstractions of satisfiable instances of a hidden SMT formula. We present a DPLL($\mathcal{T}$)-like framework to build a solver for this problem. We then prove that the methodology of Symbolic Execution also fits our framework. Based on these ideas, we build two QIF analysis tools: the first one employs CBMC, a bounded model checker for ANSI C, and the second one is built on top of Symbolic Pathfinder, a Symbolic Execution tool for Java. For experiment, we quantify information leakage of programs in the Linux kernel, and analyse a tax program in Java taken from the European project HATS.

3.20 Quantitative Distributed Data Usage Control

Alexander Pretschner (TU München, DE)

Joint work of Pretschner, Alexander; Lovat, Enrico; Oudinet, Johan

Distributed data usage control is about what happens to data once it is given away (“delete after 30 days;” “notify me if data is forwarded;” “copy at most twice”). In the past, we have considered the problem in terms of policies, enforcement and guarantees from two perspectives: (a) In order to protect data, it is necessary to distinguish between content (a song by Elvis called "Love me Tender") and representations of that content (song.mp3; song.wav, etc.). This requires data flow-tracking concepts and capabilities in data usage control frameworks. (b) These representations exist at different layers of abstraction: a picture downloaded from the internet exists as pixmap (window manager), as element in the browser-created DOM tree (application), and as cache file (operating system). This requires the data flow tracking capabilities to transcend the single layers to which they are deployed. In distributed systems, it has turned out that another system can be seen as another set of abstraction layers, thus generalizing the basic model. Demo videos of this work available at http://www22.in.tum.de/forschung/distributed-usage-control/.

In this talk, we present recent work on extending our approach to not only protecting entire data items but possibly also fractions of data items. This allows us to specify and enforce policies such as “not more than 20% of the data may leave the system”, evidently leading to interesting questions concerning the interpretation of “20%”, and if the structure of data items cannot be exploited. We present a model, its instantiation to the operating system layer, and first experimental results.

3.21 Constraint Solving Based on Horn Clauses for Verifying Information Flow Properties

Andrey Rybalchenko (TU München, DE)

Joint work of Gleissenthall, Klaus; Köpf, Boris

We present a constraint generation and solving methods that can be used as building blocks for automatic verification of information flow properties of programs.

3.22 Quantifying Opacity

Mathieu Sassolas (Université Libre de Bruxelles, BE)

Joint work of Bérard, Béatrice; Mullins, John; Sassolas, Mathieu
Main reference B. Bérard, J. Mullins, M. Sassolas, "Quantifying Opacity," arXiv:1301.6799v2 [cs.CR]. To appear in Mathematical Structures in Computer Science, Cambridge University Press.
URL http://arxiv.org/abs/1301.6799

Opacity is a general language-theoretic scheme [1] which can be instanciated into several security properties of a system. Its parameters are a predicate, given as a subset of runs of the system, and an observation function, from the set of runs into a set of observables. The predicate describes secret information in the system and, in the possibilistic setting, it is opaque if its membership cannot be inferred from observation.

In this presentation, we propose several notions of quantitative opacity for probabilistic systems, where the predicate and the observation function are seen as random variables. The distribution of these variables is based on the distribution of the potentially infinite set of runs, and therefore is not given in an extensive form.

Our aim is to measure (i) the probability of opacity leakage relative to these random variables and (ii) the level of uncertainty about membership of the predicate inferred from observation. We show how these measures extend possibilistic opacity, we give algorithms to compute them for regular secrets and observations, and we apply these computations on the classical example of Crowds protocol.

As part of ongoing work, we also study approximate computation of these measures and the non-deterministic setting.

This talk is based on joint work with Béatrice Bérard and John Mullins [2].

References

1 Bryans, J.W., Koutny, M., Mazaré, L., Ryan, P.Y.A.: Opacity generalised to transition systems. Intl. Jour. of Information Security **7**(6) (2008) 421–435

2 Bérard, B., Mullins, J., Sassolas, M.: Quantifying opacity. Mathematical Structures in Computer Science (To appear)

3.23 Thermodynamic aspects of confidentiality: timing channels in Brownian computers

Fabrizio Smeraldi (Queen Mary University of London, GB)

Joint work of Malacaria, Pasquale; Smeraldi, Fabrizio

Timing channels are contingent both on the specific computation and on the architecture of the system - notably, the presence of a clock signal. Brownian computers are unclocked devices that operate through the combined action of thermal agitation and a weak driving potential (think computing with DNA). Under these conditions variations of the entropy of the computer due to, for instance, logically irreversible computations generally affect the dynamics of the system.

As a consequence, an entirely new category of timing channels emerges that allows discriminating between computations requiring the same number of steps on the basis of their different degree of irreversibility.

The talk is based on a forthcoming paper [1].

References

1 P. Malacaria, F. Smeraldi.: Thermodynamic Aspects of Confidentiality. Information and Computation (To appear).

3.24 Channels and Composition Refinement

Geoffrey Smith (Florida Int. Univ. – Miami, US)

Joint work of Espinoza, Barbara

Given channels C1 and C2 from a set X of secret inputs, it may be that C1 is equivalent to C2 followed by some post-processing; that is, C1 can be factored into the cascade of C2 and C3 for some channel C3. In this case we say that C1 is composition refined by C2. Composition refinement coincides with partition refinement in the Lattice of Information in the case when C1 and C2 are deterministic channels, but composition refinement is meaningful for probabilistic channels as well. In this talk, I discuss some current work on the mathematical structure of channels under the composition refinement relation, considering in particular the case when channels C1 and C2 are composition equivalent, meaning that each composition refines the other. I show that composition refinement is a partial order up to semantic equivalence, where channels are semantically equivalent if they are equal as maps from priors to hyper- distributions. I also mention some connections to Quantitative Information Flow.

The talk is based on joint work with Barbara Espinoza.

3.25 On Complexity of Verifying Quantitative Information Flow

Tachio Terauchi (Nagoya University, JP)

Joint work of Terauchi, Tachio; Yasuoka; Hirotoshi
Main reference H. Yasuoka, T. Terauchi, "On Bounding Problems of Quantitative Information Flow," Journal of Computer Security, 19(6):1029–1082, 2011, IOS Press.
URL http://dx.doi.org/10.3233/JCS-2011-0437

We present results on hardness of precisely checking and inferring a program's quantitative information flow (QIF). The results are presented from two perspectives:

1. verification theoretic view,
2. complexity theoretic view.

In 1.), we classify various QIF problems into program verification problem classes such as safety and liveness (and their recently-proposed extensions such as hypersafety and hyperliveness). This reveals that different QIF definitions, such as min entropy and Shannon entropy, often exhibit different hardness for some QIF problems. In 2.), we give complexity theoretic hardness results, focusing on boolean programs. The results uphold the classification given in 1.), and also show close connection of some of the QIF problems to counting problems.

3.26 Bayesian inference to evaluate information leakage in complex scenarios: case study mix-networks

Carmela Troncoso (Gradiant – Vigo, ES)

This work casts the trace analysis of anonymity systems, and in particular mix networks, in the context of Bayesian inference. A generative probabilistic model of mix network architectures is presented, that incorporates a number of attack techniques in the trace analysis literature. We use the model to build an Markov Chain Monte Carlo inference engine, that calculates the probabilities of who is talking to whom given an observation of network traces. We provide a thorough evaluation of its correctness and performance, and confirm that mix networks with realistic parameters are secure. This approach enables us to apply established information theoretic anonymity metrics on complex mix networks, and extract information from anonymised traces optimally. This work is further explained and evaluated in [1, 2].

References

1 C. Troncoso, G. Danezis, *The Bayesian Analysis of Mix Networks,* Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS 2009), E. Al-Shaer, S. Jha, A. D. Keromytis, Eds., ACM, pp. 369-379, 2009.

2 C. Troncoso, *Design and analysis methods for privacy technologies,* PhD thesis, Katholieke Universiteit Leuven. C. Diaz, B. Preneel (advisors), 189+17 pages, 2011.

3.27 Computer-Aided Proofs of Differential Privacy

Santiago Zanella Beguelin (Microsoft Research UK – Cambridge, GB)

Joint work of Zanella Beguelin, Santiago; Barthe, Gilles; Danezis, George; Grégoire, Benjamin; George; Kunz, César
URL http://easycrypt.gforge.inria.fr

Differential privacy permits the disclosure of noisy statistics of sensible data without compromising the privacy of individuals. Proving that a program that uses standard sanitization mechanisms guarantees differential privacy is relatively easy. However, proving that an arbitrary probabilistic program guarantees differential privacy is an error-prone task that calls for a principled approach and tool support. In this talk, I will show how the novel relational program logic of Barthe et al. [2] can be used to reason about differential privacy as well as its approximate and computational relaxations. Moreover, I will report on the implementation of a proof system for this logic into the EasyCrypt interactive prover [1], a tool that can be used to verify the security of cryptographic primitives in the computational model. The resulting integrated proof assistant allows users to verify privacy guarantees of general probabilistic programs that use cryptography, under computational assumptions. Finally, I will illustrate the use of this framework to verify that a two-party protocol for computing Hamming distance between bit-vectors yields two-sided privacy guarantees.

References

1 G. Barthe, B. Grégoire, S. Heraud, and S. Zanella Béguelin, "Computer-aided security proofs for the working cryptographer," in *Advances in Cryptology – CRYPTO 2011*, ser. Lecture Notes in Computer Science, vol. 6841. Heidelberg: Springer, 2011, pp. 71–90.

2 G. Barthe, B. Köpf, F. Olmedo, and S. Zanella Béguelin, "Probabilistic relational reasoning for differential privacy," in *39th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL 2012*. New York: ACM, 2012, pp. 97–110.

Participants

- Alessandro Aldini
Univ. of Urbino, IT
- Mario Alvim
University of Pennsylvania, US
- Anindya Banerjee
IMDEA Software Institute, ES
- Béatrice Bérard
UPMC, Lab. LIP6 – Paris, FR
- Arnar Birgisson
Chalmers UT – Göteborg, SE
- Michele Boreale
University of Firenze, IT
- Kostas Chatzikokolakis
Ecole Polytechnique – Palaiseau, FR
- Tom Chothia
University of Birmingham, GB
- David Clark
University College London, GB
- Jorge Cuellar
Siemens – München, DE
- Alessandra Di Pierro
Univ. degli Studi di Verona, IT
- Ehab ElSalamouny
Ecole Polytechnique – Palaiseau, FR
- Sardaouna Hamadou
Ecole Polytechnique – Palaiseau, FR
- Holger Hermanns
Universität des Saarlandes, DE
- Michael Hicks
University of Maryland – College Park, US
- Sebastian Hunt
City University – London, GB
- Daniel Kifer
Penn State University – University Park, US
- Boris Köpf
IMDEA Software Institute, ES
- Matteo Maffei
Universität des Saarlandes, DE
- Pasquale Malacaria
Queen Mary University of London, GB
- Fabio Martinelli
CNR – Pisa, IT
- Michael W. Mislove
Tulane University, US
- C. Carroll Morgan
UNSW – Sydney, AU
- John Mullins
Ecole Polytechnique – Montreal, CA
- Gethin Norman
University of Glasgow, GB
- Martin Ochoa
Siemens – München, DE
- Catuscia Palamidessi
Ecole Polytechnique – Palaiseau, FR
- Quoc Sang Phan
Queen Mary University of London, GB
- Alexander Pretschner
TU München, DE
- Andrey Rybalchenko
TU München, DE
- Mathieu Sassolas
Université Libre de Bruxelles, BE
- Vladimiro Sassone
University of Southampton, GB
- Fabrizio Smeraldi
Queen Mary University of London, GB
- Geoffrey Smith
Florida Int. Univ. – Miami, US
- Marco Stronati
Ecole Polytechnique – Palaiseau, FR
- Tachio Terauchi
Nagoya University, JP
- Carmela Troncoso
Gradiant – Vigo, ES
- Herbert Wiklicky
Imperial College London, GB
- Santiago Zanella Beguelin
Microsoft Research UK – Cambridge, GB

Report from Dagstuhl Seminar 12482

Analysis of Security APIs

Edited by

Mike Bond[1], Riccardo Focardi[2], Sibylle Fröschle[3], and Graham Steel[4]

1 **University of Cambridge, GB,** mike.bond@cl.cam.ac.uk
2 **Università Ca' Foscari di Venezia, IT,** focardi@dsi.unive.it
3 **Universität Oldenburg, DE**
4 **INRIA, FR,** graham.steel@inria.fr

Abstract

This report documents the programme and the outcomes of Dagstuhl Seminar 12482 "Analysis of Security APIs". Abstracts from the talks give a snapshot of current research in the field, while reports on the discussions give a roadmap for future research in the area.

Seminar 26.–28. November, 2012 – www.dagstuhl.de/12482
1998 ACM Subject Classification K.6.m Miscellaneous, Security
Keywords and phrases Security APIs, cryptography, key management, formal methods, security protocols
Digital Object Identifier 10.4230/DagRep.2.11.155

1 Executive Summary

Mike Bond
Riccardo Focardi
Sibylle Fröschle
Graham Steel

This report documents the programme and outcomes of Dagstuhl Seminar 12482 "Analysis of Security APIs". The seminar brought together 32 participants from academia and industry in Europe and the USA. It featured a joint session with the concurrent seminar on quantitative security analysis (which included the keynote talk), a breakout session with demonstrations of software and practical classes, a discussion of the most important open problems in the field and a collection of talks spanning the breadth of the field from theoretical models to applications.

Research Context and Goals of the Seminar

A security API is an Application Program Interface that allows untrusted code to access sensitive resources in a secure way. It is the interface between processes running with different levels of trust. Examples of security APIs include the interface between the tamper-resistant chip on a smartcard (trusted) and the code running on the client application (untrusted), the interface between a cryptographic Hardware Security Module (or HSM, trusted) and the host machine (untrusted), and web service APIs (an interface between a server, trusted by the service provides, and the rest of the Internet).

Analysis of Security APIs, *Dagstuhl Reports*, Vol. 2, Issue 11, pp. 155–168
Editors: Mike Bond, Riccardo Focardi, Sibylle Fröschle, and Graham Steel
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The crucial aspect of a security API is that it is designed to enforce a policy, i.e. no matter what sequence of commands in the interface are called, and no matter what the parameters, certain security properties should continue to hold. This means that if the less trusted code turns out to be malicious (or just faulty), the carefully designed API should prevent compromise of critical data. Designing such an interface is extremely tricky and error prone, and over the last ten years, serious vulnerabilities in the security APIs deployed in HSMs in the ATM (cash machine) network and in commodity security devices like smartcards and USB keys have come to light.

A number of formal methods researchers have turned their attention to security APIs over the last five years. While significant advances have been made and notable results achieved, such as the discovery of several new attacks, the process of specifying and verifying the security policy for such APIs still lacks both satisfactory foundations and efficient algorithms. At the same time, the security API paradigm is being proposed as a solution for more and more applications, from social networks to smartphones, with more complicated and less well understood security goals.

The objective of the seminar was to bring together researchers in academia and industry around the topic of security APIs and their analysis. There were three main aims:

1. To address the shortcomings of current API analysis techniques as applied to the relatively well explored domains of cryptographic key management and HSMs, in particular in their ability to deal with global mutable state and their models of cryptography.
2. To identify the security API requirements arising from the new generation of applications, in mobile device applications and web services, and map out the research problems that need to be solved in order that formal API analysis can be applied here.
3. To find ways to include the process and results of formal API analysis into the standards and certification procedures.

Some progress was made on all these points in the talks and the discussions late into the evening that followed in the conducive environment of Schloss Dagstuhl. On the first point, several talks presented models specifically aimed at modelling state in a more satisfactory way, and we had a tutorial on the verification methods used in program analysis. Several new application areas for API analysis were presented, including car to car communication and password protection. Some highly enlightening talks on the standards process helped to improve understanding of the problem, if not providing steps to an easy solution. The variety of open problems identified (see summary below) shows that this is a vibrant area of computer security research with much promise for the future.

2 Table of Contents

3 Keynote: Ross Anderson – Security evolution: interaction of economics and APIs

Ross Anderson opened the second day of the seminar in a joint session with the quantitative security analysis workshop[1]. The talk was an exciting journey through the evolution of IT security from both technical and socio-economic perspectives. Economics matters: failures are inversely proportional to how much is invested into guarding and fixing a system; lack of security is often considered an external, independent factor we might need to address, just like environmental pollution. Moreover, security is often an obstacle: to win the market a new platform has little security so that development is easier; once the market is captured security can be faced. Ross finally illustrated a few impressive "non-success stories" on payment and banking systems and pointed out how the APIs often are the places of interest, i.e. "where the rubber hits the road".

4 Overview of Talks

The technical programme included demos, hand-on sessions, perspectives, work in progress and new ideas as well as conventional research talks.

4.1 Security in Car2X Communication

Daniel Angermeier (Fraunhofer AISEC – München, D E)

Joint work of Angermeier, Daniel; C2C Communication Consortium

Car2X communication promises improvements in modern traffic, as cooperative driving might help avoid dangerous situations. Furthermore, C2X communication aims to achieve improved traffic efficiency. However, these potential advantages are opposed by risks caused especially by attacks on Intelligent Transport Systems (ITS) trying to abuse these new features. Therefore, security plays a major role in ITSs to reach the afore mentioned goals and to avert threats caused by attackers. In my talk, I will highlight a few aspects of security in C2X communication, which focus on fulfilling the special requirements in C2X communication, like e.g., privacy of ITS users or proof of trustworthiness of received messages.

[1] http://www.dagstuhl.de/12481

4.2 Demo of Tookan: Tool for Cryptoki Analysis

Romain Bardou (INRIA – Paris, FR)

Joint work of Bardou, Romain; Focardi, Riccardo; Steel, Graham
Main reference M. Bortolozzo, M. Centenaro, R. Focardi, G. Steel, "Attacking and Fixing PKCS#11 Security Tokens," in Proc. of the 17th ACM Conf. on Computer and Communications Security (CCS'10), pp. 260–269, ACM, 2010.
URL http://dx.doi.org/10.1145/1866307.1866337
URL http://tookan.gforge.inria.fr/

This is a demonstration of Tookan, a tool which automatically finds attacks on cryptographic devices. Tookan reverse engineers the behavior of PKCS#11 tokens. It learns the preconditions of each command to build a model of the token. It then runs a model-checking analysis on the model to try and find a sequence of commands leading to an invalid state. Tookan can be used for penetration testing, but it can also be used to compare device configurations, or to help develop a safe cryptographic device.

4.3 Tokenisation – pseudo-security and compliance engineering

Mike Bond (University of Cambridge, GB)

This talk describes the challenges in designing a security solution whose main goal is to satisfy compliance requirements for international financial standards such as the PCI Data Security Standard (PCI DSS). Together with highly prescriptive internal security standards there is often very little room to design an elegant or efficient solution and this proves very costly for organisations that must be compliant. Sometimes the compliance rules even fly in the face of security concerns, or are contradictory/ill-defined. For instance, tokenisation is defined by some to be a deterministic substitution mechanism which is not an algorithmic function, yet a look-up table is indeed a function. Sometimes the challenge becomes to modify a solution design to avoid falling foul of compliance rules without introducing significant vulnerability, and sometimes the challenge is to actively frustrate trivial data flow analysis of a solution such as is used by many auditors who simply follow the flows of keys and data and then make broad prescriptions about the wisdom or otherwise of a scheme. The talk proposes some schemes for 'audit-resistant cryptography', and shows their application to practical problem solving in an environment tainted by conflicting security and compliance requirements.

4.4 Revoke and Let Live: A Secure Key Revocation API for Cryptographic Devices

Veronique Cortier (CNRS – Nancy, FR)

Main reference V. Cortier, G. Steel, C. Wiedling, "Revoke and Let Live: A Secure Key Revocation API for Cryptographic Devices," in Proc. of the 19th ACM Conf. on Computer and Communications Security (CCS'12), pp. 918–928, ACM, 2012.
URL http://dx.doi.org/10.1145/2382196.2382293

While extensive research addresses the problem of establishing session keys through cryptographic protocols, relatively little work has appeared addressing the problem of revocation and update of long term keys. We present an API for symmetric key management on embedded devices that supports revocation and prove security properties design in the symbolic model of cryptography. Our API supports two modes of revocation: a passive mode where keys have an expiration time, and an active mode where revocation messages are sent to devices. For the first we show that once enough time has elapsed after the compromise of a key, the system returns to a secure state, i.e. the API is robust against attempts by the attacker to use a compromised key to compromise other keys or keep the compromised key alive past its validity time. For the second we show that once revocation messages have been received the system is immediately in a secure state. Notable features of our designs are that all secret values on the device are revocable, and the device returns to a functionally equivalent state after revocation is complete.

4.5 Hands-on tutorial on Padding Oracle Attacks

Riccardo Focardi (Università Ca' Foscari di Venezia, IT)

Joint work of Bardou, Romain; Focardi, Riccardo; Kawamoto, Yusuke; Simionato, Lorenzo; Steel, Graham; Tsay, Joe-Kai
Main reference R. Bardou, R. Focardi, Y. Kawamoto, L. Simionato, G. Steel, J. Tsay, "Efficient Padding Oracle Attacks on Cryptographic Hardware," in Proc. of 32nd Annual Conf. on Advances in Cryptology (CRYPTO'12), LNCS, Vol. 7417, pp. 608–625, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-32009-5_36
URL http://secgroup.ext.dsi.unive.it/wp-content/uploads/2012/11/Practical-Padding-Oracle-Attacks-on-RSA.html

We revise attacks on the RSA cipher based on side-channels that leak partial information about the plaintext. We show how to compute a plaintext when only its parity is leaked. We then describe PKCS#1 v1.5 padding for RSA and we show that the simple leakage of padding errors is enough to recover the whole plaintext, even when it is unpadded or padded under another scheme. This vulnerability is well-known since 1998 but the flawed PKCS#1 v1.5 padding is still broadly in use. We discuss recent optimizations of this padding oracle attack that make it effective on commercially available cryptographic devices. We illustrate through many examples and fragments of code. This tutorial is based on the paper appeared in Hakin9 – Defend Yourself! Hands-on Cryptography, September 2012, available at http://secgroup.ext.dsi.unive.it/wp-content/uploads/2012/11/Practical-Padding-Oracle-Attacks-on-RSA.html

4.6 Challenges in Security API Verification

Sibylle Froeschle (OFFIS – Oldenburg, DE)

Main reference S. Froeschle, "Causality in security protocols and security APIs: foundations and practical verification," Habilitation Thesis, University of Oldenburg, 2012.
URL https://vhome.offis.de/sibyllef/habil.pdf

In this talk we pinpoint key challenges in security API verification and suggest possible solutions and research directions. Among the challenges we discuss are the problem of scale and several general aspects such as how to specify security APIs and what to verify about them. A central theme will be how to deal with the key metadata that governs how a key entity is managed by the API. The talk is based on a comparative study of PKCS#11 and IBM's CCA, two widely deployed key management APIs. A detailed discussion can be found in Chapter I.4 of [1]

References

1 Sibylle Fröschle. Causality in security protocols and security APIs: foundations and practical verification. Habilitation thesis, University of Oldenburg, 2012.

4.7 Universally Composable Key-Management

Steve Kremer (INRIA Grand Est – Nancy, FR)

Main reference St. Kremer, R. Künnemann, G. Steel, "Universally Composable Key-Management," Cryptology ePrint Archive: Report 2012/189, 2012.
URL http://eprint.iacr.org/2012/189

We present the first universally composable key management functionality, formalized in the GNUC framework by Hofheinz and Shoup. It allows the enforcement of a wide range of security policies and can be extended by diverse key usage operations with no need to repeat the security proof. We illustrate its use by proving an implementation of a security token secure with respect to arbitrary key-usage operations and explore a proof technique that allows the storage of cryptographic keys externally, a novel development in simulation-based security frameworks.

4.8 Formal Security Analysis Results for the Yubikey and YubiHSM

Robert Kuennemann (CNRS, INRIA, FR)

The Yubikey is a small hardware device designed to authenticate a user against network-based services. Despite its widespread adoption (over a million devices have been shipped by Yubico to more than 20 000 customers including Google and Microsoft), the Yubikey protocols have received relatively little security analysis in the academic literature. In the first part of this paper, we give a formal model for the operation of the Yubikeyone-time password (OTP) protocol. We prove security properties of the protocol for an unbounded number of fresh

OTPs using a protocol analysis tool, tamarin. In the second part of the talk, we analyze the security of the protocol with respect to an adversary that has temporary access to the authentication server. To address this scenario, Yubico offers a small Hardware Security Module (HSM) called the YubiHSM, intended to protect keys even in the event of server compromise. We show if the same YubiHSM configuration is used both to set up Yubikeys and run the authentication protocol, then there is inevitably an attack that leaks all of the keys to the attacker. Our discovery of this attack lead to a Yubico security advisory in February 2012. For the case where separate servers are used for the two tasks, we give a configuration for which we can show using the same verification tool that if an adversary that can compromise the server running the Yubikey-protocol, but not the server used to set up new Yubikeys, then he cannot obtain the keys used to produce one-time passwords.

4.9 A Framework for the Cryptographic Verification of Java-like Programs

Ralf Kuesters (Universität Trier, DE)

Joint work of Küsters, Ralf; Truderung, Tomasz; Graf, Jürgen
Main reference R. Küsters, T. Truderung, J. Graf, "A Framework for the Cryptographic Verification of Java-like Programs," in Proc. of the 25th IEEE Computer Security Foundations Symp. (CSF'12), pp. 198–212, IEEE CS, 2012.
URL http://dx.doi.org/10.1109/CSF.2012.9

We consider the problem of establishing cryptographic guarantees—in particular, computational indistinguishability—for Java or Java-like programs that use cryptography. For this purpose, we propose a general framework that enables existing program analysis tools that can check (standard) non-interference properties of Java programs to establish cryptographic security guarantees, even if the tools a priori cannot deal with cryptography. The approach that we take is new and combines techniques from program analysis and simulation-based security. Our framework is stated and proved for a Java-like language that comprises a rich fragment of Java. The general idea of our approach should, however, be applicable also to other practical programming languages. As a proof of concept, we use an automatic program analysis tool for checking non-interference properties of Java programs, namely the tool Joana, in order to establish computational indistinguishability for a Java program that involves clients sending encrypted messages over a network, controlled by an active adversary, to a server. The approach may also be applicable for checking security properties of Java programs that use security APIs.

4.10 Lazy Mobile Intruders

Sebastian Moedersheim (Technical University of Denmark, DK)

Joint work of Moedersheim, Sebastian; Nielson, Hanne Riis; Nielson, Flemming
Main reference S. Moedersheim, F. Nielson, H.R. Nielson, "Lazy Mobile Intruders," in Proc. of the 2nd Int'l Conf. on Principles of Security and Trust (POST'13), LNCS, Vol. 7796, pp. 147–166, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-36830-1_8
URL http://www.imm.dtu.dk/ samo/mobile.pdf

We present a new technique for analyzing platforms that execute potentially malicious code, such as web-browsers, mobile phones, or virtualized infrastructures. Rather than analyzing given code, we ask what code an intruder could create to break a security goal of the platform. To avoid searching the infinite space of programs that the intruder could come up with (given some initial knowledge) we adapt the lazy intruder technique from protocol verification: the code is initially just a process variable that is getting instantiated in a demand-driven way during its execution. We also take into account that by communication, the malicious code can learn new information that it can use in subsequent operations, or that we may have several pieces of malicious code that can exchange information if they "meet". To formalize both the platform and the malicious code we use the mobile ambient calculus, since it provides a small, abstract formalism that models the essence of mobile code.

References

1 Sebastian Mödersheim, Flemming Nielson, Hanne Riis Nielson. *Lazy Mobile Intruders*. In Proceedings of POST 2013, Springer LNCS, 2013. Extended version available as IMM-TR-2012-13 at www.imm.dtu.dk/~samo.

4.11 Temporal Information Flow

Markus N. Rabe (Universität des Saarlandes, DE)

Joint work of Dimitrova, Rayna; Finkbeiner, Bernd; Kovács, Máté; Rabe, Markus N.; Seidl, Helmut
Main reference R. Dimitrova, B. Finkbeiner, M. Kovács, M.N. Rabe, H. Seidl, "Model Checking Information Flow in Reactive Systems," in Proc. of the 13th Int'l Conf. on Verification, Model Checking, and Abstract Interpretation (VMCAI'12), LNCS, Vol. 7148, pp. 169–185, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-27940-9_12
URL http://www.react.uni-saarland.de/publications/DFKRS12.html

There is a great number of different security properties that are discussed in the various security communities, but they are defined on different semantic models and are thus difficult to compare. Further, there is currently little interface to other specifications that concern the safety and lifeness aspects of a system. I will present recent results on how to integrate secrecy properties into temporal logics by introducing a new modal operator. Besides providing a common framework for many security properties, this allows to precisely specify when and under which conditions a variable has to be kept secret and also until when the secrecy needs to be maintained. I also give an overview of the complexity results and an efficient fragment that is suitable for checking large models.

4.12 APIs and Cryptography

Phillip Rogaway (University of California – Davis, US)

In this talk I discussed, in turn: (1) how an API can inform a cryptographic definition (example: the notion of online indistinguishability from [Rogaway, Wooding, Zhang 2012]); (2) how an API can inform the design of a cryptographic algorithm (example: incremental encryption and OCB3 [Krovetz, Rogaway 2011]); and (3) how cryptographic expertise can (maybe poorly) inform the design of an API (example the GCS-API [Rogaway 1994] that I developed at IBM).

4.13 Automated Reasoning on Data Structures

Viorica Sofronie-Stokkermans (Universität Koblenz-Landau, DE)

We present a class of theories for which the problem of checking satisfiability of ground formulae is decidable. Examples are theories of data structures (fragments of theories of arrays or pointers), as well as extensions of certain classes of base theories with functions which satisfy certain recursion and homomorphism properties. We present the applications of these ideas in verification and possibly also in cryptography.

4.14 MAC in the Box

Graham Steel (CNRS, INRIA, FR)

We propose to construct a formally verified open source security device for calculating messages authentication codes (MACs). The application we have in mind is the storage of user passwords on a web server. Typically, these are stored as salted hashes of the password and some other diversifiers (such as username). Unfortunately, password files are often leaked after server compromise, and computing power is sufficiently affordable to allow brute force offline cracking of the passwords. To prevent this we propose to calculate a keyed hash (HMAC) of passwords. All HMACs will be calculated in a separate hardware device (the MAC in the Box, or MITB) where the key is stored. The API of the device will allow calculation and verification of HMACs but no commands will give access to the key. In the event of server compromise, the attacker's ability to crack passwords is limited by the throughput of the MITB. After the compromise is discovered and the attacker is ejected, the password file is of no use to him, since he has no access to the HMAC key. We anticipate that such a simple device could be formally verified to a low level. In combination with a low cost and open source design the MITB will be an attractive best-practice option for website administrators.

4.15 Verification of a Trusted Virtual Security Module

Ronald Toegl (TU Graz, AT)

Joint work of Toegl, Ronald; Reimair, Florian; Pirker, Martin
Main reference R. Toegl, F. Reimair, M. Pirker, "Waltzing the Bear, or: A Trusted Virtual Security Module," in Proc. of the 9th European PKI Workshop: Research and Applications (EUROPKI 2012), LNCS, Springer, in press.
URL https://online.tugraz.at/tug_online/voe_main2.getVollText?pDocumentNr=286273&pCurrPk=67562

Cryptographic key material needs to be protected. Currently, this is achieved by either pure software based solutions or by more expensive dedicated hardware security modules. We present a practical architecture to project the security provided by the Trusted Platform Module and Intel Trusted eXecution Technology on a virtual security module. Our approach uses commodity personal computer hardware to offer integrity protection and strong isolation to a security module which implements a compact security API that has been fully verified. Performance results suggest that our approach offers an attractive balance between speed, security and cost.

5 Discussion

In the final session, participants were asked to describe one thing they had learnt during the seminar and one important topic for future research (either to be conducted by themselves or by others). Here we highlight some of the most interesting suggestions:

5.1 What I learned

- The auditor as an adversary. Mike Bond's talk on the socio-technical side of standards and certification prompted several comments. It is clear that what formal researchers analyse is not always relevant for practice – e.g. because of standards. Also one can see that compliance is sometimes damaging usability and security. In particular, "certification as compliance" is seen as eroding to the value of the certification process.
- Even small APIs are useful and present interesting design and verification challenges
- Formal researchers were frequently struck by the range of applications of HSMs, and the practicalities of their use (e.g. in a mixed estate of heterogeneous configurations).
- Being an area that attracts theoreticians but also practitioners, many researchers found it interesting to consider the security economics angle of API analysis, as outlined by Ross Anderson in his talk.
- Practitioners were generally pleased to see that tools were on the way. In particular, competition between approaches seems healthy. There is no need to work on just one approach for the moment.
- Many attendees found the tension between theory and practice interesting

5.2 Future research topics

Many new security APIs ripe for analysis were suggested, including:

- Low level APIs of crypto devices
- OS (e.g. SE linux security modules).

- Geographic security APIs – e.g. vehicular.
- Heterogeneous networks of APIs – e.g. different HSMs with different APIs in networks.

Other topics included:

- Improved design of APIs around Crypto considerations
- Languages for human interpretation of APIs and policy. Human behavioural studies that could lead to comprehensible security policies.
- "Better than Dolev Yao" models (i.e. more cryptographic detail)
- A formally verified, open source HSM
- Concurrency and its effects on security – inside devices/drivers/applications
- Privacy properties of security APIs – e.g. in V2X
- A simple common language for APIs.

6 Conclusion

The field of security API analysis is in rude health. The seminar was over subscribed and the participation by attendees enthusiastic. As well as consolidating well known subjects in the area, the seminar identified new research directions in foundations and applications. The next few years should be an exciting time for research in this area.

Participants

- Pedro Adao
IST – TU of Lisbon, PT
- Ross Anderson
University of Cambridge, GB
- Daniel Angermeier
Fraunhofer AISEC –
München, DE
- David R. Aspinall
University of Edinburgh, GB
- Romain Bardou
INRIA – Paris, FR
- Mike Bond
University of Cambridge, GB
- Veronique Cortier
CNRS – Nancy, FR
- Marion Daubignard
Direction Générale de
l'Armement, FR
- Stéphanie Delaune
CNRS, ENS – Cachan, FR
- Riccardo Focardi
Univ. Ca' Foscari di Venezia, IT
- Sibylle Fröschle
OFFIS – Oldenburg, DE
- Steve Kremer
INRIA Grand Est – Nancy, FR
- Robert Künnemann
CNRS, ENS – Cachan, FR
- Ralf Küsters
Universität Trier, DE
- Flaminia L. Luccio
Univ. Ca' Foscari Venezia, IT
- Matteo Maffei
Universität des Saarlandes, DE
- Sebastian Mödersheim
Technical Univ. of Denmark, DK
- Benjamin Morin
ANSSI -Paris, FR
- Andreas Philipp
Utimaco Safeware AG, DE
- Markus N. Rabe
Universität des Saarlandes, DE
- Phillip Rogaway
Univ. of California – Davis, US
- Mark D. Ryan
University of Birmingham, GB
- Stefanie Schlegel
OFFIS – Oldenburg, DE
- Jörg-Cornelius Schneider
Deutsche Bank – Eschborn, DE
- Laurent Simon
University of Cambridge, GB
- Viorica Sofronie-Stokkermans
Universität Koblenz-Landau;
MPI für Informatik, Saarbrücken
- Marco Squarcina
Univ. Ca' Foscari di Venezia, IT
- Graham Steel
CNRS, ENS – Cachan, FR
- Petr Svenda
Masaryk University, CZ
- Susan Thompson
MasterCard Worldwide,
Warrington, GB
- Frank Thunig
Utimaco Safeware AG, DE
- Ronald Toegl
TU Graz, AT