



DAGSTUHL REPORTS

Volume 5, Issue 4, April 2015

Assuring Resilience, Security and Privacy for Flexible Networked Systems and Organisations (Dagstuhl Seminar 15151) <i>David Hutchison, Klara Nahrstedt, Marcus Schöller, Indra Spiecker gen. Döhmman, and Markus Tauber</i>	1
Machine Learning with Interdependent and Non-identically Distributed Data (Dagstuhl Seminar 15152) <i>Trevor Darrell, Marius Kloft, Massimiliano Pontil, Gunnar Rätsch, and Erik Rodner</i>	18
Advanced Stencil-Code Engineering (Dagstuhl Seminar 15161) <i>Christian Lengauer, Matthias Bolten, Robert D. Falgout, and Olaf Schenk</i>	56
Software and Systems Traceability for Safety-Critical Projects (Dagstuhl Seminar 15162) <i>Jane Cleland-Huang, Sanjai Rayadurgam, Patrick Mäder, and Wilhelm Schäfer</i> ..	76
Theory and Practice of SAT Solving (Dagstuhl Seminar 15171) <i>Armin Biere, Vijay Ganesh, Martin Grohe, Jakob Nordström, and Ryan Williams</i>	98
Challenges and Trends in Probabilistic Programming (Dagstuhl Seminar 15181) <i>Gilles Barthe, Andrew D. Gordon, Joost-Pieter Katoen, and Annabelle McIver</i> ...	123
Qualification of Formal Methods Tools (Dagstuhl Seminar 15182) <i>Darren Cofer, Gerwin Klein, Konrad Slind, and Virginie Wiels</i>	142

ISSN 2192-5283

Published online and open access by

Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at <http://www.dagstuhl.de/dagpub/2192-5283>

Publication date

December, 2015

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

License

This work is licensed under a Creative Commons Attribution 3.0 DE license (CC BY 3.0 DE).



In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work under the following conditions, without impairing or restricting the authors' moral rights:

- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

Aims and Scope

The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.

In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

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

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

Editorial Board

- Bernd Becker
- Stephan Diehl
- Hans Hagen
- Hannes Hartenstein
- Oliver Kohlbacher
- Stephan Merz
- Bernhard Mitschang
- Bernhard Nebel
- Bernt Schiele
- Nicole Schweikardt
- Raimund Seidel (*Editor-in-Chief*)
- Arjen P. de Vries
- Michael Waidner
- Reinhard Wilhelm

Editorial Office

Marc Herbstritt (*Managing Editor*)
Jutka Gasiorowski (*Editorial Assistance*)
Thomas Schillo (*Technical Assistance*)

Contact

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

Digital Object Identifier: 10.4230/DagRep.5.4.i

Report from Dagstuhl Seminar 15151

Assuring Resilience, Security and Privacy for Flexible Networked Systems and Organisations

Edited by

David Hutchison¹, Klara Nahrstedt², Marcus Schöller³,
Indra Spiecker gen. Döhmann⁴, and Markus Tauber⁵

1 Lancaster University, GB, d.hutchison@lancaster.ac.uk

2 University of Illinois at Urbana-Champaign, US, klara@illinois.edu

3 Hochschule Reutlingen, DE, marcus.schoeller@reutlingen-university.de

4 Goethe-Universität Frankfurt, DE, spiecker@jur.uni-frankfurt.de

5 AIT Austrian Institute of Technology – AT, markus.tauber@ait.ac.at

Abstract

Dagstuhl Seminar 15151 entitled “Assuring Resilience, Security and Privacy for Flexible Networked Systems and Organisations” brought together researchers from different disciplines in order to establish a research agenda for making future services in our increasingly connected world more resilient and secure, as well as addressing privacy. The participants came from a range of disciplines covering the techno-legal domain, resilience and systems security, and socio-technical concerns. The use case domains that were discussed during the Seminar covered the Internet of Things (IoT) as well as Cloud-based applications in which flexible service composition is a crucial element. From a starting point covering the “big picture”, the legal viewpoint, the technical viewpoint, and the organisational viewpoint, we derived initial research questions in small groups, and the questions and issues arising were then consolidated and refined. The groups discussed the issues in depth and have produced the report and the research agenda contained here.

Seminar April 7–10, 2015 – <http://www.dagstuhl.de/15151>

1998 ACM Subject Classification C.2 Computer-communication networks, J.4 Social and behavioural sciences, K.4 Computers and society, K.5 Legal aspects of computing

Keywords and phrases Resilience, security, privacy, legal aspects, networked systems, organisations, society

Digital Object Identifier 10.4230/DagRep.5.4.1

Edited in cooperation with Simon Oechsner (NEC, simon.oechsner@neclab.eu)

1 Executive Summary

David Hutchison

Klara Nahrstedt

Marcus Schöller

Indra Spiecker gen. Döhmann

Markus Tauber

License © Creative Commons BY 3.0 Unported license

© David Hutchison, Klara Nahrstedt, Marcus Schöller, Indra Spiecker gen. Döhmann, and Markus Tauber

This report documents the programme and the outcomes of Dagstuhl Seminar 15151 on “Assuring Resilience, Security and Privacy for Flexible Networked Systems and Organisations”.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Assuring Resilience, Security and Privacy for Flexible Networked Systems and Organisations, *Dagstuhl Reports*, Vol. 5, Issue 4, pp. 1–17

Editors: David Hutchison, Klara Nahrstedt, Marcus Schöller, Indra Spiecker gen. Döhmann, and Markus Tauber



DAGSTUHL
REPORTS

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

The main objective of the Seminar was to bring together researchers from different disciplines in order to establish a research agenda for securing services-to-come in our increasingly connected world. The backgrounds and interests of the participants included i) techno-legal, ii) resilience and systems security, and iii) socio-technical topics. The use case domains that were discussed covered the Internet of Things (IoT) as well as Cloud-based applications in which flexible service composition is paramount. We started the seminar using four introductory talks covering respectively the “big picture”, the legal viewpoint, the technical viewpoint, and the organisational viewpoint. From this beginning, we derived initial research questions in small groups, and these questions and issues arising were then consolidated and refined into the resulting material that is presented below.

The opening speakers were the following:

- Helmut Leopold, Head of the Digital Safety and Security Department at the Austrian Institute of Technology, who presented the “big picture”, i.e. where our connected world is heading;
- Burkhard Schafer, Professor of Computational Legal Theory at the University of Edinburgh, who presented his viewpoint on legal challenges within our ever interconnected society;
- Thilo Ewald from Microsoft Deutschland GmbH, who explained his viewpoint on the organisational challenges in today’s world;
- Marcus Brunner, Head of Standardization in the strategy and innovation department of Swisscom, presented his viewpoint on technological developments in designing and building flexible networked systems.

From this starting point we derived initial research questions in small groups. The organising team reviewed intermediate results and re-balanced groups and most significantly identified the core questions to work on. The groups were between 4 and 6 people at any time, and a good balance was maintained across the representatives of legal, organisational and technological experts and between the groups. The resulting questions and issues were:

1. How to enable Resilience, by design, of composable flexible systems [1]?
2. What is the role of law in supporting resilience, privacy [2] and security?
3. Traceability of (personal and non-personal) data in service provision?
4. How can we improve the perception of assurance [3], privacy, security and resilience for the end-user?
5. What constitutes a security problem?
6. How to deal with unforeseen new context of usage?

These questions were crucial, in that they formed the basis for the bulk of group discussions throughout the second and third days of the Seminar. Therefore, the organisers took great care – and a great deal of time during the first evening – formulating these questions, together with the related issues. At the start of the second day, these questions and issues were presented to the groups, who were invited to comment on them. The groups were invited to add their own interpretation, and to identify additional issues during their discussions. During the subsequent periods – broken up by refreshments and lunch – the organisers checked that the groups appeared to be productive and harmonious (which on both counts they turned out to be). Each group was asked to record the essence of their discussions, and conclusions, and to pass these to the organisers by the end of the Seminar. Every group did some additional work after the Seminar, and the report assembled here reflects the hard work of the participants as well as the organisers, during the Seminar itself and in the days that followed.

References

- 1 James P. G. Sterbenz, David Hutchison, Egemen K. Çetinkaya, Abdul Jabbar, Justin P. Rohrer, Marcus Schöller, and Paul Smith. Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines. *Comput. Netw.*, 54(8):1245–1265, June 2010.
- 2 Burkhard Schafer. All changed, changed utterly? *Datenschutz und Datensicherheit – DuD*, 35(9):634–638, 2011.
- 3 Aleksandar Hudic, Markus Tauber, Thomas Lorunser, Maria Krotsiani, George Spanoudakis, Andreas Mauthe, and Edgar R. Weipl. A multi-layer and multitenant cloud assurance evaluation methodology. In *Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on*, pages 386–393. IEEE, 2014.

2 Table of Contents

Executive Summary

David Hutchison, Klara Nahrstedt, Marcus Schöller, Indra Spiecker gen. Döhmann, and Markus Tauber 1

The report

How we ran the Seminar 5
Introductory Talks 5
Research Questions 6

Participants 17

3 The report

3.1 How we ran the Seminar

Opening talks took place on the first morning; on the first afternoon we ran pre-selected groups to produce candidate questions, which were consolidated by the organisers during the first evening.

Initial group setup, at least one organizer was part of the groups (where the organiser responsible for each group is identified in square brackets):

1. Brunner Balaban Alshawish Mauthe Tsudik [Tauber]
2. Leopold Raabe Fischer Sterbenz Varga [Hutchison]
3. Ewald Lyles Sorge Stiller Weippl [Spiecker]
4. Pallas Kadobayashi Kirby Smith Schaeffer-Filho [Schoeller]
5. Schafer Bhatti Delsing Bless Dan [Oechsner¹]

We readjusted the groups for the second day, as follows:

1. Brunner Raabe Alshawish Mauthe [Tauber]
2. Leopold Balaban Fischer Sterbenz Varga [Hutchison]
3. Ewald Sorge Stiller Tsudik [Spiecker]
4. Pallas Kadobayashi Kirby Smith Schaeffer-Filho [Schoeller]
5. Schafer Bhatti Delsing Bless Dan [Oechsner]

There followed in-depth discussion of research questions that the groups themselves chose freely based on the above list (the organisers checked that there was sufficient disparity across groups).

At the end of each session, there was a presentation of discussion outcomes from each group.

3.2 Introductory Talks

The introductory talks and other presentations can be found in the appendix:

- Big Picture: Helmut Leopold
As head of the department Digital Safety and Security at the Austrian institute of technology Helmut Leopold supervises research agendas in multiple fields. He presented his view on the general directions and major trends in research and society. Those included ICT Trends in the „after-broadband century”, the Security problem, the Shift in user behaviour and the IT industry problem.
- Legal Viewpoint: Burkhard Schafer
Burkard Schafer is Professor of Computational Legal Theory at the University of Edinburg and co-director of the Joseph Bell Centre for Legal Reasoning and Forensic Statistics. As such he operates on the intersection between law, science and computer technology. He presented his view point on gaps in this particular intersection and requirements regarding legal definitions. He also pointed out that security research ethics need to be established to allow research a structured way of publishing their empirical findings.

¹ Simon Oechsner supported the organisation team in the absence of Klara Nahrstedt.

- **Organisational Viewpoint: Thilo Ewald**
Thilo Ewald is Microsoft's Cloud Delivery Executive for Germany. As such he has an overview of the technologies to come and customers requirements and anticipations. He presented the challenges of deploying scalable global office solutions and how security relates to scalability and connectivity.
- **Technology Viewpoint: Marcus Brunner**
Marcus Brunner is responsible at Swisscom's Strategy and Innovation unit for standardisation issues which he is leading. He presented technology and customer expectation issues. Focusing on telco provider issues and concluding that customers see security often as part of the telco provision.

Additional presentations were done on:

- **Connecting Legacy Systems**
Jerker Delsing is Professor at Lulea University of Technology. He talked about the problem of connecting legacy technology to the internet of everything and supported this with solutions [1] from the arrowhead project².
- **Address-space, Routing and Mobility**
Saleem Bhatti is Professor at the University of St Andrews. He talked about problems related addressing in the internet of everything and supported this with solutions [2] from the ilnp initiative³.

3.3 Research Questions

The research questions produced after the opening presentations, which formed the basis for the core group sessions in the Seminar, are reproduced here:

1. How to enable Resilience, by design, of composable flexible systems (new architectural models)?
 - How to overcome complexity?
 - Responsibility
 - Risk-management and assessment
2. What is the role of law in supporting resilience, privacy and security (technology regulation)?
 - How to enforce law – e.g. automated algorithmic law enforcement?
3. Traceability of (personal and non-personal) data in service provision
 - Anonymization and de-anonymization of data
4. How can we improve the perception of assurance, privacy, security and resilience for the end-user?
 - Can we build an economic model for security and resilience?
 - The role of trust in assuring security
5. What constitutes a security problem (model, define, measure, . . .)?
 - Do we need an ethical framework for handling findings from security research?
6. How to deal with unforeseen new context of usage?
 - Shifting responsibility for data

² <http://www.arrowhead.eu/>

³ <http://ilnp.cs.st-andrews.ac.uk/>

Each group chose freely from amongst these questions – with the associated issues – to discuss in depth the topics that they find important and can recommend be included in a research agenda.

A summary of the findings for each research question / topic is given below. This is followed in the Appendix by a presentation of the report that each group produced, providing a correlation of the outcome of the Seminar.

3.3.1 Question 1

How to enable Resilience, by design, of composable flexible systems (new architectural models)?

- How to overcome complexity?
- Responsibility
- Risk-management and assessment

The first question to be discussed was how to specify resilience at the enterprise level and how to map this into the system layers and mechanisms.

The group recognized that this is similar to the Quality of Service (QoS) mapping issue that was a subject of considerable research during the 1990s. Specification of desired qualities has to be done at the enterprise (or application level) in a form or in a language that is understood by the end-user, using for example the so-called Olympic levels of Gold, Silver or Bronze, or alternatively using some QoS classes that implicitly encapsulate the desired QoS properties – such as interactive applications as opposed to file transfer (where in the former case, delay and jitter must be minimized, but in the latter, these are less important). In order to communicate the QoS specification into the network, a mapping has to be made from the high-level statement into the corresponding network parameters or metrics (thus, for example, delay, jitter, packet loss and so on). Research into the specification of resilience and the mapping into appropriate metrics is ongoing (notably by some participants in this Dagstuhl seminar), and builds on research that was carried out in the EU Framework Future Internet ResumeNet project. One of the key issues is what metrics to use at the service and topology levels, though much more work has been done on the latter than the former.

We therefore agreed that a service level agreement (SLA) driven system design (for composable systems and services) is appropriate, though this raises related issues of the relationship to policy, regulation, and the law. This is further elaborated in the next paragraph. A study of trust boundaries in composed and multi-level systems was also agreed to be important, along with the related policy and legal implications.

The formulation of SLAs needs to be adapted in composed systems, as the probability of a failure grows with the amount of involved parties (as it is intended in composed systems). From a legal viewpoint, it is mandatory to distinguish between external and internal relationships within composed systems. Whereas the external one encompasses the contractual relationship between an end user and the composed system provider, the internal deals with the contractual relationship of the composed system provider and (other) providers of systems s/he himself uses to perform the duties owed to clients. In this situation, the provider of a composed system has different obligations. First, s/he is responsible for the systems s/he himself offers to end users (ensuring that the services being relied on work together as intended). She or he is, in turn, indirectly also responsible for the (contractual) performance of suppliers and their subsystems. These relationships define implicitly a trust boundary already. This boundary can be made explicit with the help of contracts, SLAs, and

technical interfaces that clearly define the owed performance and functionality, and therefore the sphere of responsibility.

However, composed (and in particular virtualized) systems can still cause difficulties with correctly attributing responsibility and liability. Although from a legal viewpoint responsibility lies with the one who acts in a negligent manner in the way that he fails to exercise reasonable care, the complexity of the system itself may make it difficult to pinpoint the cause of a failure (i.e., is it a subsystem or the composition of such systems), and therefore identify the responsible legal party. Monitoring and/or recording systems may be helpful in this situation to assess what actually happened, and permit the party who bears the burden of proof to get an inside view that can support further legal prosecution.

Therefore, the SLAs have to take the higher probability of liability due to the increased number of acting parties into account. Still, it is not obvious that even a well-formulated SLA can procure a non-liability if it is technically not possible to determine what actually happened.

A different question that arose was the following: can we structure (or architect) systems to create boundaries or interfaces that act as trust boundaries, or at least as clear functional or perhaps ownership boundaries. The latter should be relatively easy to achieve, though this immediately reverts to the issue of how to create and agree levels of trust between owners or operators of parts of the infrastructure. A related issue is that of interface abstractions and tussles between entities that are – or that may be – unwilling to exchange information or to agree on trust levels. This is clearly a potential impediment to the successful construction of resilient systems. We simply agreed to add this to our resilient systems research agenda.

Two further, related questions are (i) in what ways are composed systems able to be structured to reduce complexity, where components are not necessarily fully described or understood, and (ii) how to model and understand, and subsequently assure, the resilience of (composed) interconnected and/or interdependent networks. The first of these is studied by complex systems researchers, which was not represented at this Dagstuhl seminar, while the second has recently become the subject of considerable interest amongst graph theorists amongst others, and is recognized by the resilience researchers participating in the seminar as being one of the most important topics for us to study because of the evident interdependence between various real-world (critical infrastructure) networks such as telecommunications, electricity distribution and public transport (for example).

Related to this is that we must ensure that resilience mechanisms do not make systems more fragile, even though we may have made them more complex, and also in introducing them we will very likely have increased the attack surface for the very systems that we are trying to protect.

We moved on to the issue of safety-critical systems, for which there is inevitably different thinking about resilience because of the societal importance of the systems in question, such as aviation, railway transport and roads networks. For these systems, the publication of information about incidents and liability to risks is considered essential and in the public interest. For other networks that can be considered critical (though not safety-critical) such as financial, government, corporate or telecommunication networks, for example, there is much less interest in discussing their resilience, and this is unfortunate when it is increasingly obvious that society really depends on these systems, and they should attract considerably more attention by owners, operators and also researchers.

A key research question, one that is of interest to some participants in the seminar, is understanding and modelling the roles of humans in (composed) systems; this includes how to assess risks, and how to assure resilience of systems in which humans are a constituent

part. Previous research has been conducted in the fields of Human Computer Interface (HCI) and Computer Supported Cooperative Work (CSCW), some time ago, and a current imperative is to study and include this prior research into the resilience research agenda.

We discussed the prospect of autonomic operation in critical systems that need to be made resilient: can removing the human in the loop make safer systems? The range of such systems we discussed included telecommunications operations, aviation, and the driverless car. We observed that in some of these systems, human on the loop (flying on automatic pilot, for example) is already much used. How well would this translate to 'driverless' cars? Clearly there would be significant implications and legal responsibilities and liabilities, which would need to be closely embedded along with discussions about the technical viability about autonomic operation.

And in such composed (especially virtualised) systems there would be difficulties attributing liability (or responsibility), even when activities have been monitored or recorded, following any incidents.

For the goal of assessment it is necessary to find new approaches that can deal with the additional complexity of composed systems, i.e., taking into account new interactions that had not been included in the design and implementation of the individual components. Designing components or systems in the face of uncertainty with respect to their usage context or environment can cause unnecessary complexity in their code to provide the desired level of resilience. This additional complexity caused by uncertain component contexts may weaken the reliability and trust of the overall system, since more possibilities for programming errors exist. Consequently, providing resilience for flexible and composable systems (FCS) is challenging. Approaches like model checking need to be examined whether they can feasibly be applied to this new environment, with an eye towards the on-demand and cost-efficient assessment of the properties of composed systems. Here, also the question of responsibility for these checks and assessments arises, i.e., how much responsibility (and liability) the original components' designers have and how much has to be borne by the composer or end-user. It may be possible to learn more in this respect from other disciplines that have faced/are facing similar issues.

It might be of interest as well to investigate how much a component designer can introduce in terms of mechanisms that are context-sensitive, i.e., change the behaviour of the component in different contexts to maintain SPR. Here, a basic trade-off between high levels of security, privacy and resilience (SPR) on the one hand and a high degree of flexibility/composability on the other needs to be evaluated, as well. One potential solution may be the regular software update of components, which may be necessary or at least desirable for fixing security issues anyway. Such updates may open the possibility to let the component interact in new contexts, but given the longevity of some devices (e.g., sensors built into houses) it is not very likely that vendors provide development and software updates for their products over such long periods.

3.3.2 Question 2

What is the role of law in supporting resilience, privacy [3] and security?

- Does technology regulation play a part?
- How to enforce law – e.g. automated algorithmic law enforcement?

Trying to clarify the potential contribution of law to privacy, security and resilience, one immediately encounters fundamental questions about the function of law in a society, as well as the relation between different approaches to regulation. Entities affected by the law,

particularly companies, expect regulation to be clear, foreseeable, and to provide certainty to enable planning of future decisions. How can these goals be achieved? There is a continuum of norms, from formal laws passed by legislators to standards agreed upon in a technical community. In general, technical standards can react to new developments faster, and given that they are mostly developed by scientists and engineers, they are easier to understand for that same target group. There is sometimes frustration in that community because they seem to be forgotten in the law-making process, so should their voice be considered more when drafting new legislation? This may sometimes improve the quality of the legal texts, but this is difficult to do in a democratic process, which requires elected politicians to be in charge. Neutral advice, of course, would be helpful, but completely neutral experts without their own agenda exist in theory only. The cultural background plays a role as well, as law enforcement alone cannot ensure legal compliance if the law itself is not considered acceptable in society.

One source of frustration when engineers deal with legal texts is the lack of concrete guidelines, which would ideally use precise numbers and thresholds. Unfortunately, this is rarely feasible, as such thresholds would be arbitrary, as seen in examples like a threshold scale from which aerial images in Google Earth are considered to contain personal information. German telecommunications regulation is an example for the inclusion of a concrete regulation model put into law; we doubt whether its interpretation by the Bundesnetzagentur actually follows the spirit of the law, though.

Despite these problems, can the law still play a role in improving security and privacy? We see attempts in Germany (IT Security Act) and the US (particularly in the health sector). Data breach disclosure requirements, for example, can serve as an incentive to improve security, without trying to go into too many details. Previous attempts of regulation in the IT sector have led to unintended results, though, in the broader context, the Oracle vs. Google case is a good example. As an example of infosec regulation that did not work, in the 1980s and 1990s, the US government tried to restrict dissemination of knowledge of public-key encryption and of strong symmetric cryptography, primarily via export control. That led to controversy and “interesting times”; but it is probably not controversial to say that no one’s aims were achieved. (Steve Levy’s book *Crypto* is a good summary of this story, but other references exist as well.)

The US NIST’s work in establishing cryptographic standards provides both positive and negative examples of effective infosec regulation. One negative example is the process by which NIST and NSA transformed IBM’s “Lucifer” algorithm into the Data Encryption Standard (DES); a prevailing belief was that NSA had installed a backdoor. In response, NIST used an open and public competition to select DES’s replacement as AES, which consequently had broader acceptance. However, some subsequent actions on the SHA-3 competition raised concerns of backdoors again (as did Snowden’s revelations about backdoors in some elliptic curve PRNGs).

A final problem to be considered in information security regulation is the conflict between different jurisdictions, both on the level of federal states and between nations. The concept of discovery, for example, which is used in the U.S., seems rather scary for European lawyers. The above-mentioned export regulation of cryptography is another example.

We also discussed the question how law and technology relate and how they contribute to privacy.

On one hand, law defines a set of rules that are supposed to determine what (among others) technology is allowed to do; on the other hand, we have the impression that legal norms seem to lag behind technical developments. This has led to norms being ignored, and national authorities have often failed to enforce them. In recent years, however, we observe a changing attitude of regulators and courts in Europe.

Common law and civil law jurisdictions have different approaches to cope with technical change. In common law, the focus is on precedents, while civil law uses more abstract norms, trying to cover future cases already in statute (though still requiring a neutral entity, i.e., a court, for arbitration). We assume that the latter approach is better suited to deal with innovation, such as the big data paradigm. European legislation protects personal data, making anonymization a core concept to enable law-compliant data processing|in the big data context, de-anonymization is often possible due to an unforeseeable amount of additional information that can be linked with the original data collection. This leads to the issue when to consider a certain anonymization procedure (such as the addition of noise) as sufficient [4].

In addition, it raises the question whether the current data protection legislation in Europe needs to be adapted to benefit from the advantages of research based on big data. One approach could be to regulate procedure (for data processing) instead of result. In practice, globalization has caused problems for legislation and law enforcement, increasing complexity and enabling circumvention; yet, it does not imply that law is powerless per se. Past experience has shown that the EU has been able to enforce European law even against the interests of global players like Microsoft.

Concerning resilience, we have discussed the impact of (de-)centralization; decentralized systems can increase resilience, but under certain circumstances, the opposite can be true. In networks, users can often change the behaviour of individual nodes, thus causing an impact on the overall system's behaviour that cannot be foreseen or controlled by the respective user or a central entity. The example of the smart grid shows that lack of resilience in IT systems can have real-world consequences. It also illustrates how privacy, security and resilience relate to each other's mechanisms improving privacy, for example, may hinder the detection of attacks, and cryptographic processing may enable DoS attacks due to high processing load.

The group has also covered the topic of trust in IT systems; under which circumstances does one need trust, and can we talk of trust if we are certain of the functionality? Snowden's revelations have, in some cases, shaken the confidence in some beliefs about IT security, thus increasing the relevance of this question once more.

Additional questions/issues which evolved in post-seminar-discussions: What are values law should encompass? What goods should law protect in the field of security and resilience, e.g. ownership and/or personality? Can law and regulatory powers be used to equalize potential market failure and/or differences in power and strength?

3.3.3 Question 3

Traceability of (personal and non-personal) data in service provision?

- Which legal and technical aspects to consider?
- Anonymization and de-anonymization of data?

Users of current and future applications and services (e.g. Gmail, Facebook, Body-Sensors, Smart-homes, etc.) must be assured that the data defining their identity remains protected (and being purpose bound). This requires a new legal definition of data ownership, because the state of the art treats data as tradable goods in a classic sense. Treating this kind of data in this sense is not appropriate since it pertains to a person and defines (directly or indirectly) their identity beyond the transaction period. Hence, it directly links to human rights like the right for privacy and in the case of misuse can ultimately violate the constitutional requirement of keeping the dignity of man sacrosanct. But personal-data still has value that should be exploitable (within a given framework). This value can be tangible (e.g.

expressed through monetary transactions) or intangible (e.g. social standing and personal reputation). In order to facilitate this, a digital “market place” is required that guarantees transaction transparency, awareness and control of personal data when being handled and exploited. Hence personal data cannot be “owned” in the traditional sense but can only be allowed to use within a well-defined legal, political, social and commercial context. In order to achieve this we believe that new standards as regulatory mechanisms are required, which should be legitimated through a democratic process. Research will have to establish how the new concepts of “market place” and “usage rights” can be realized within a technical framework that enables traceability of data and its usage and protects it from misuse (such as unauthorised trading) but still allowing for the realization of new market concepts.

This approach can be extend to systems which would be required to interoperate with each other in dynamic use cases where some may be operated by third parties on behalf of a user which may change over time. A market place approach as above described would already allow management of identity of data. An extension regarding the mapping of resources and services to such identities may allow for management of responsibilities in dynamic unforeseen situations. Future scenarios may include multiple Google-smart-home instances, interoperating with some e-health application.

Additional questions/issues which evolved in post-seminar-discussions: What are the different concepts of identity in law and in technical sciences? What are incentives of the involved parties to accept legal restrictions and how can they be created?

3.3.4 Question 4

How can we improve the perception of assurance [5], privacy, security and resilience for the end-user?

- Can we build an economic model for security and resilience?
- How to specify the role of trust in assuring security?

To ensure security and resilience of (distributed) flexible and composable systems (FCS), in an ideal world, security researchers would be able to test in advance every piece of software or application for potential problems it may encounter, which would allow them to mitigate the problems through adequately modifying the system design. This being overly optimistic, we could still hope for identifying problems as they occur and give adequate warnings to stakeholders, who would then take all and only those actions required to mitigate the risk/damage. Reality often looks different, with information about risks either not available or not distributed timely, or warnings exaggerated and resulting in panic rather than measured response. While not a new problem, FCS is likely to increase the seriousness of these issues.

Regarding FCS, two aspects are particularly challenging: how to measure or assess security, privacy and resilience (SPR), and how to communicate the results of this assessment to the end user of the systems (or other relevant stakeholders). Since assessment is discussed separately, we here focus more on the communication and transparent aspect. Regardless, in both of these aspects one of the main challenges is the fact that, by definition, flexible systems are operated in varying and thus often unforeseeable contexts, e.g., in new service compositions, in new environments, or for new purposes. This characteristic makes it particularly difficult to analyse such systems a priori, and in principle necessitates mechanisms that differ from existing approaches designed for static systems. It may be possible to learn more in this respect from other disciplines that have faced/are facing similar issues. Research in medical drugs could be one such comparator: it is one thing to establish if a drug, taken on its own,

is harmful for a patient. However, it is impossible to foresee what other medication(s) a patient may be taking in addition, outside the confines of a tightly controlled medical trial. There are however mechanisms that with varying success try to address this issue, from voluntary or mandatory reporting mechanisms if an incompatibility has been experienced, to the information leaflet that informs the patient on what other drugs she should avoid. Some of this is underpinned by a risk management strategy. For more serious drugs, pharmacists will ask a set of questions on other medication taken before releasing it to the buyer, with less risky drugs it is left to the patient etc. A possible research question should look at the success or failure of these approaches in cognate fields, and explore to what degree the analogy to FCSs is valid.

The necessity for regulation of such assessment and information mechanisms will depend on a classification of the severity and impact of issues and dangers. Critical, high impact systems with the potential for severe damage (e.g., energy utility systems) should be treated differently than systems that might only have individual, personal effects (although a cumulative effect might be taken into account if these minor damages occur for a large set of end-users). However, social expectations on what counts as “acceptable risk” are changing as rapidly as the technologies themselves. A few years ago, there was no Facebook. Today, even a very short temporary outage leaves people at the very least in emotional distress – some however face real difficulties, as they rely on Facebook to sign into other, more crucial, systems. Here, social practices can cause a loss in resilience that is difficult to foresee or counteract in a FCS environment.

The question of propagating the results to end-users is somewhat dependent on the possibilities for and outcomes of assessment mechanisms. It relates to the concept of trust and how the perception of trust can be enhanced. One basic question is how detailed information about the SPR levels is to be made available to end-users (e.g., a ‘five star’ system similar to car safety tests, combined abstract and more detailed information like existing energy efficiency classes, or even more detailed reports).

Another is how security breaches or similar critical events are communicated to the parties affected. In the past, this happened largely on an ad-hoc basis, with little or patchy regulation. A conventional virus checker for instance will give some warnings instantaneously, but because this requires automated detection and notification, the untrained user is given relatively little in information on “what to do now”. By contrast, a security breach in a credit card company will be communicated, if at all, to customers through the established media with a degree of time delay (or individually, by email or similar channel), but with the advantage of careful advice tailor made to the situation (“it was a minor breach, it is unlikely that your credit card details were released, however you should change your password for this site. If you have further concerns, cancel the card or call our helpline. . .”).

For a world of FCS, we should rethink if these channels of communication are still adequate (if they ever were). With automated bug reporting for instance, to which party should a report be sent if the problem emerged from the ad-hoc interaction between several machines and programs? If we surround ourselves with gadgets and get too many warnings, is there a danger of the “boy who shouted wolf” syndrome, so that we get desensitized and stop taking appropriate actions? Should several machines negotiate with each other which one has to inform a user of an issue, so as not to cause information overload (e.g.: my fridge, thermostat, washing machine and car decide between each other which one to alert me if a problem they all experience has a single cause).

Other connected questions are which level of detail is to be mandatory for selected systems (again linked to the risk classes described above), and to which degree this information needs

to be tailored to the expertise level of the recipient and be made easily accessible. Related to this is investigating the need to educate users about the significance of security information, particularly the 'digital natives', the younger people growing up being used to IT technology and maybe not sensitive to security and privacy issues.

Interesting aspects here might also be the evaluation of the feasibility of wisdom of the crowds approaches (e.g., a futures market for possible attacks, similar to the "futures market in terrorist attacks" that the US government briefly contemplated), or to what degree the existing biological or psychological models for trust are of value in the context of FCS – can we find design solutions that use our evolved mechanisms to ascribe or revoke trust and optimize them?

Apart from legally mandatory information disclosure about critical system characteristics, some of the tasks of informing the end users might also be taken over by market mechanisms. For example, a service or system provider offering more transparency for each customer may be more trustworthy and therefore get more customers. If a sufficient market landscape of service and system operators exists, each with its own approach to transparency in addition to the mandatory one, user preference might lead to the automatic establishment of standards of information. Here relevant research questions should address the drivers and obstacles for an efficient market in FCS – Intellectual property rights (de-jure monopolies) or standard setting for instance could prevent the emergence of an efficient "market in security".

Another market-related incentive for service providers to provide information about its security is if a model for insuring such services can be developed where having higher levels of security and transparency, at least from the viewpoint of the insurance companies, directly results in benefits for the insured provider. In such a scenario, being able to prove that a system is secure would bring direct economic benefits, e.g., being able to show operation without security incidents or usage of more secure systems and receiving a decrease in premiums. On the other hand, insurance companies might also be able to exert pressure by refusing coverage for services where no sufficient transparency is provided. Reliable mechanisms for transparency and auditing are again necessary and useful to this end. An open issue in this context are how feasible such a model is considering the possibility of presently unknown risks such as large scale vulnerabilities and exploits discovered in the future for any service.

Learning again from risk management in other fields could be of benefit. In the regulation of financial services, strict rules exist on what (and when) investors need to be informed, e.g., in the form of a "profit warning". At the same time, individual investors in the UK e.g. will get from their independent financial advisor or investment broker a mandatory "risk profile" that tells them what products are suitable for them given their willingness (and resilience to) certain risks. Could something similar, in machine readable and transportable format be relevant for FCS?

These approaches rely on a final decision on trustworthiness by a human and thus might be of limited use in the important field of application of machine-to-machine communication or automated composition of systems. It is to be seen whether in environments where there are no human intervention or decision other mechanisms are necessary, making trust understandable and utilizable for machines, or if maybe the concept of trust can only be applied for systems where humans are involved. In a world where machines or devices may automatically close contracts, automated trust assessment and delegation may be necessary. One possibility could be developing a machine behaviour code describing acceptable machine behaviours related to an automatically closed contract. This in combination with code breaching detection mechanisms would be an additional component helping to increase SPR.

However, trust and reputation systems tend to become complex and may serve as an attractive attack target instead of attacking other security mechanisms (e.g., attacking a TLS-secured communication may be easier by using illegitimate certificates). On the other hand, removing humans and thus the potential for human error from the loop might also have benefits if reliable automated systems can be designed. In any case, the legal implications of and responsibility attribution in a pure machine-to-machine environment also need to be explored.

It should be noted that an increased importance of assessment and information systems and reliance on the information thus propagated also increases the attractiveness of such meta-systems themselves for attacks. Care needs to be taken that no new avenues of attack are created by mechanisms that themselves are used to assure end users of the resilience, security or privacy of other systems. For instance, a competitor might try to exploit automated fault monitoring and reporting systems by flooding them with faked incidences of faults in a competitor's product (similar to the manipulation of reviews and ratings that we already find on e-commerce and recommendation sites) – there is also a question of how the law should proscribe, if at all, this type of behaviour and impose (criminal law?) sanctions.

Additional questions/issues which evolved in post-seminar-discussions: What role can other actors not directly involved, e.g. insurances, play? Can trust be developed in purely non-human interaction?

3.3.5 Question 5

What constitutes a security problem?

- How to model, define, measure, ... security problems?
- Do we need an ethical framework for handling findings from security research?

It is necessary to inquire into the currently existing and maybe insufficient research infrastructure and culture regarding especially security research. Intellectual property law and data protection law for instance have been accused of hampering necessary security research. While EU data protection law recognizes a “research exemption”, there is at least some evidence that this provision is badly understood and insufficient in allaying the fears of administrators in university ethics committees. At the very least, the question should be asked if this type of provision that was tailor made for medical research “fits” the practice of security research in FCS. One possibility would be to clarify (or create) “research exceptions” in copyright and data protection law. A potentially more appealing solution would be to restrict these exceptions to a special class of “bona fide security researchers” with additional exemption from legal prosecution for their type of research. Here, lessons could be learned from the very different way the EU and the US regulate journalism as a profession that is also (partially) exempted from data protection rules. At the same time, a code of conduct needs to be established for this research, in particular for rules for the publication of research results that might lead to an increased risk due to the disclosure of security flaws. Another example would be the necessity to report findings independently of their potential to attract attention, i.e., reports of negative results in the sense of not reporting any flaws should be treated the same way as reports about vulnerabilities deemed more ‘interesting’ for the public. Such a research culture would have as its goal a more thorough and independent research that is thus also perceived as more trustworthy by the general public and by the subjects of investigation. Learning again from the experience with safety research in medicine, it might be worth exploring if there ought to be a “notification scheme” for certain types of research projects and repositories for research findings, to prevent the “publication bias” inherent

in traditional research. In the US, the department for Homeland Security recently made available huge datasets for security research in IT infrastructures through the PREDICT repository. A promising research project would be to evaluate the suitability of this database for research in FCS, and setting up, if needed, a similar system for FCS. The PREDICT approach to data privacy would need to be analysed to ensure its acceptability within an EU setting.

3.3.6 Question 6

How to deal with unforeseen new context of usage?

- What legal and technical dimensions are involved?
- What to expect when shifting responsibility for data?

Even though “How to deal with data usage in a new context” was identified as a stand-alone research topic for a research agenda, we believe that the contribution to topic 3 “Traceability of (personal and non-personal) data in service provision (anonymisation and de-anonymisation of data)” addresses the topic perfectly well.

Additional Material

Original presentations including introductory talks and supporting presentations can be found at these URLs:

- <http://materials.dagstuhl.de/files/15/15151/15151.SWM2.Slides1.ppt>
- <http://materials.dagstuhl.de/files/15/15151/15151.SWM3.Slides.pptx>
- <http://materials.dagstuhl.de/files/15/15151/15151.SWM4.Slides.pptx>
- <http://materials.dagstuhl.de/files/15/15151/15151.SWM5.Slides.pptx>
- <http://materials.dagstuhl.de/files/15/15151/15151.JerkerDelsing.Slides.pdf>
- <http://materials.dagstuhl.de/files/15/15151/15151.SaleemBhatti.Slides.pdf>

References

- 1 Rumén Kyusakov, Pablo Punal Pereira, Jens Eliasson, and Jerker Delsing. Exip: a framework for embedded web development. *ACM Transactions on the Web (TWEB)*, 8(4):23, 2014.
- 2 S. N. Bhatti, D. Phoomikiatissak, and R. J. Atkinson. Fast, Secure Failover for IP. In *MILCOM 2014 – 33rd IEEE Military Communications Conf.*, Oct 2014.
- 3 Burkhard Schafer. All changed, changed utterly? *Datenschutz und Datensicherheit – DuD*, 35(9):634–638, 2011.
- 4 Burkhard Schafer, Judith Rauhofer, Zbigniew Kwecka, and William Buchanan. “I am Spartacus”: privacy enhancing technologies, collaborative obfuscation and privacy as a public good. *Artificial Intelligence and Law*, 22:113–139, 2014.
- 5 Aleksandar Hudic, Markus Tauber, Thomas Lorunser, Maria Krotsiani, George Spanoudakis, Andreas Mauthe, and Edgar R. Weippl. A multi-layer and multitenant cloud assurance evaluation methodology. In *Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on*, pages 386–393. IEEE, 2014.

Participants

- Ali Alshawish
Universität Passau, DE
- Silvia Balaban
KIT – Karlsruher Institut für
Technologie, DE
- Saleem Bhatti
University of St. Andrews, GB
- Roland Bless
KIT – Karlsruher Institut für
Technologie, DE
- Marcus Brunner
Swisscom AG – Bern, CH
- György Dan
KTH Royal Institute of
Technology, SE
- Jerker Delsing
Luleå Univ. of Technology, SE
- Thilo Ewald
Microsoft Deutschland GmbH –
Unterschleissheim, DE
- Andreas Fischer
Universität Passau, DE
- David Hutchison
Lancaster University, GB
- Youki Kadobayashi
Nara Institute of Science and
Technology, JP
- Graham Kirby
University of St. Andrews, GB
- Helmut Leopold
AIT Austrian Institute of
Technology – Wien, AT
- Andreas Mauthe
Lancaster University, GB
- Simon Oechsner
NEC Laboratories Europe –
Heidelberg, DE
- Frank Pallas
KIT – Karlsruher Institut für
Technologie, DE
- Oliver Raabe
KIT – Karlsruher Institut für
Technologie, DE
- Alberto Egon Schaeffer-Filho
Federal University of Rio Grande
do Sul, BR
- Burkhard Schafer
University of Edinburgh, GB
- Marcus Schöller
Hochschule Reutlingen, DE
- Sean W. Smith
Dartmouth College –
Hanover, US
- Christoph Sorge
Universität des Saarlandes –
Saarbrücken, DE
- Indra Spiecker gen. Döhmann
Goethe-Univ. Frankfurt, DE
- James P. G. Sterbenz
University of Kansas, US
- Burkhard Stiller
Universität Zürich, CH
- Markus Tauber
AIT Austrian Institute of
Technology – Wien, AT
- Gene Tsudik
Univ. of California – Irvine, US
- Pal Varga
Budapest University of
Technology & Economics, HU
- Edgar R. Weippl
Secure Business Austria
Research, AT



Machine Learning with Interdependent and Non-identically Distributed Data

Edited by

Trevor Darrell¹, Marius Kloft², Massimiliano Pontil³,
Gunnar Rätsch⁴, and Erik Rodner⁵

1 University of California - Berkeley, US, trevor@eecs.berkeley.edu

2 HU Berlin, DE, kloft@hu-berlin.de

3 University College London, GB, m.pontil@cs.ucl.ac.uk

4 Memorial Sloan-Kettering Cancer Center – New York, US

5 Friedrich Schiller University Jena, DE, erik.rodner@uni-jena.de

Abstract

One of the most common assumptions in many machine learning and data analysis tasks is that the given data points are realizations of independent and identically distributed (IID) random variables. However, this assumption is often violated, e.g., when training and test data come from different distributions (dataset bias or domain shift) or the data points are highly interdependent (e.g., when the data exhibits temporal or spatial correlations). Both scenarios are typical situations in visual recognition and computational biology. For instance, computer vision and image analysis models can be learned from object-centric internet resources, but are often rather applied to real-world scenes. In computational biology and personalized medicine, training data may be recorded at a particular hospital, but the model is applied to make predictions on data from different hospitals, where patients exhibit a different population structure. In the seminar report, we discuss, present, and explore new machine learning methods that can deal with non-i.i.d. data as well as new application scenarios.

Seminar April 7–10, 2015 – <http://www.dagstuhl.de/15152>

1998 ACM Subject Classification G.3 Probability and Statistics, I.4.8. Scene Analysis, J.3 Biology and Genetics

Keywords and phrases machine learning, computer vision, computational biology, transfer learning, domain adaptation

Digital Object Identifier 10.4230/DagRep.5.4.18

1 Executive Summary

Erik Rodner

Trevor Darrell

Marius Kloft

Massimiliano Pontil

Gunnar Rätsch

License  Creative Commons BY 3.0 Unported license

© Erik Rodner, Trevor Darrell, Marius Kloft, Massimiliano Pontil, and Gunnar Rätsch

The seminar broadly dealt with *machine learning*, the area of computer science that concerns developing computational methods using data to make accurate predictions. The classical machine learning theory is built upon the assumption of independent and identically distributed random variables. In practical applications, however, this assumption is often violated,



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Machine Learning with Interdependent and Non-identically Distributed Data, *Dagstuhl Reports*, Vol. 5, Issue 4, pp. 18–55

Editors: Trevor Darrell, Marius Kloft, Massimiliano Pontil, Gunnar Rätsch, and Erik Rodner



DAGSTUHL
REPORTS Dagstuhl Reports

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

for instance, when training and test data come from different distributions (dataset bias or domain shift) or when the data exhibits temporal or spatial correlations. In general, there are three major reasons why the assumption of independent and identically distributed data can be violated:

1. The draw of a data point influences the outcome of a subsequent draw (inter-dependencies).
2. The distribution changes at some point (non-stationarity).
3. The data is not generated by a distribution at all (adversarial).

The seminar focused on the scenarios (a) and (b). This general research direction comprises several subfields of machine learning: transfer and multi-task learning, learning with inter-dependent data, and two application fields, that is, visual recognition and computational biology. Both application areas are not only two of the main application areas for machine learning algorithms in general, but their recognition tasks are often characterized by multiple related learning problems that require transfer and multitask learning approaches. For example, in visual recognition tasks, object categories are often visually related or hierarchically organized, and tasks in computational biology are often characterized by different but related organisms and phenotypes. The problems and techniques discussed during the seminar are also important for other more general application areas, such as scientific data analysis or data-oriented decision making.

Results of the Seminar and Topics Discussed

In the following, the important research fields related to the seminar topic are introduced and we also give a short list of corresponding research questions discussed at the seminar. In contrast to other workshops and seminars often associated with larger conferences, the aim of the Dagstuhl seminar was to reflect on open issues in each of the individual research areas.

Foundations of Transfer Learning

Transfer Learning (TL) [2, 18] refers to the problem of retaining and applying the knowledge available for one or more source tasks, in order to efficiently develop an hypothesis for a new target task. Each task may contain common (domain adaptation [25, 10]) or different label sets (across category transfer). Most of the effort has been devoted to binary classification [23], while interesting practical transfer problems are often intrinsically multi-class and the number of classes can increase in time [17, 22]. Accordingly the following research questions arise:

- How to formalize knowledge transfer across multi-class tasks and provide theoretical guarantees on this setting?
- Moreover, can inter-class transfer and incremental class learning be properly integrated?
- Can learning guarantees be provided when the adaptation relies only on pre-trained source hypotheses without explicit access to the source samples, as it is often the case in real world scenarios?

Foundations of Multi-task Learning

Learning over multiple related tasks can outperform learning each task in isolation. This is the principal assertion of Multi-task learning (MTL) [3, 7, 1] and implies that the learning process may benefit from common information shared across the tasks. In the simplest case,

the transfer process is symmetric and all the tasks are considered as equally related and appropriate for joint training. Open questions in this area are:

- What happens when the condition of equally related tasks does not hold, e.g., how to avoid negative transfer?
- Moreover, can non-parametric statistics [27] be adequately integrated into the learning process to estimate and compare the distributions underlying the multiple tasks in order to learn the task similarity measure?
- Can recent semi-automatic methods, like deep learning [9] or multiple kernel learning [13, 12, 11, 4], help to get a step closer towards the complete automatization of multi-task learning, e.g., by learning the task similarity measure?
- How can insights and views of researcher be shared across domains (e.g., regarding the notation of *source task selection* in reinforcement learning)?

Foundations of Learning with Inter-dependent Data

Dependent data arises whenever there are inherent correlations in between observations. For example, this is to be expected for time series, where we would intuitively expect that instances with similar time stamps have stronger dependencies than ones that are far away in time. Another domain where dependent data occurs are spatially-indexed sequences, such as windows taken from DNA sequences. Most of the body of work on machine learning theory is on learning with i.i.d. data. Even the few analyses (e.g., [28]) allowing for “slight” violations of the assumption (mixing processes) analyze the same algorithms as in the i.i.d. case, while it should be clear that also novel algorithms are needed to most effectively adapt to rich dependency structures in the data. The following aspects have been discussed during the seminar:

- Can we develop algorithms that exploit rich dependency structures in the data?
- Do such algorithms enjoy theoretical generalization guarantees?
- Can such algorithms be phrased in a general framework in order to jointly analyze them?
- How can we appropriately measure the degree of inter-dependencies (theoretically) such that it can be also empirically estimated from data (overcoming the so-called *mixing* assumption)?
- Can theoretical bounds be obtained for more practical dependency measures than mixing?

Visual Transfer and Adaptation

Visual recognition tasks are one of the main applications for knowledge transfer and adaptation techniques. For instance, transfer learning can put to good use in the presence of visual categories with only a few number of labels, while across category transfer can help to exploit training data available for related categories to improve the recognition performance [14, 21, 20, 22]. Multi-task learning can be applied for learning multiple object detectors [30] or binary image classifiers [19] jointly, which is beneficial because visual features can be shared among categories and tasks. Another important topic is domain adaptation, which is very effective in object recognition applications [24], where the image distribution used for training (source domain) is different from the image distribution encountered during testing (target domain). This distribution shift is typically caused by a data collection bias. Sophisticated methods are needed as in general the visual domains can differ in a combination of (often unknown) factors including scene, object location and pose, viewing angle, resolution, motion blur, scene illumination, background clutter, camera characteristics, etc. Recent studies have demonstrated a significant degradation in the performance of state-of-the-art image

classifiers due to domain shift from pose changes [8], a shift from commercial to consumer video [5, 6, 10], and, more generally, training datasets biased by the way in which they were collected [29].

The following open questions have been discussed during the seminar:

- Which types of representations are suitable for transfer learning?
- How can we extend and update representations to avoid negative transfer?
- Are current adaptation and transfer learning methods efficient enough to allow for large-scale continuous visual learning and recognition?
- How can we exploit huge amounts of unlabeled data with certain dependencies to minimize supervision during learning and adaptation?
- Are deep learning methods already compensating for common domain changes in visual recognition applications?

Application Scenarios in Computational Biology

Non-i.i.d. data arises in biology, e.g., when transferring information from one organism to another or when learning from multiple organisms simultaneously [31]. A scenario where dependent data occurs is when extracting local features from genomic DNA by running a sliding window over a DNA sequence, which is a common approach to detect transcription start sites (TSS) [26]. Windows close by on the DNA strand – or even overlapping – show stronger dependencies than those far away. Another application scenario comes from statistical genetics. Many efforts in recent years focused on models to correct for population structure [16], which can arise from inter dependencies in the population under investigation. Correcting for such rich dependency structures is also a challenge in prediction problems in machine learning [15]. The seminar brought ideas together from the different fields of machine learning, statistical genetics, Bayesian probabilistic modeling, and frequentist statistics. In particular, we discussed the following open research questions:

- How can we empirically measure the degree of inter-dependencies, e.g., from a kinship matrix of patients?
- Do theoretical guarantees of algorithms (see above) break down for realistic values of “the degree of dependency”?
- What are effective prediction and learning algorithms correcting for population structure and inter-dependencies in general and can they be phrased in a general framework?
- What are adequate benchmarks to evaluate learning with non-i.i.d. data?
- How can information be transferred between organisms, taking into account the varying noise level and experimental conditions from which data are derived?
- How can non-stationarity be exploited in biological applications?
- What are promising applications of non-i.i.d. learning in the domains of bioinformatics and personalized medicine?

Conclusion

The idea of the seminar bringing together people from theory, algorithms, computer vision, and computational biology, was very successful, since many discussions and joint research questions came up that have not been anticipated in the beginning. These aspects were not completely limited to non-i.i.d. learning and also touched ubiquitous topics like learning with deeper architectures. It was the agreement of all participants that the seminar should be the beginning of an ongoing series of longer Dagstuhl seminars focused on non-i.i.d. learning.

References

- 1 Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.
- 2 Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.
- 3 Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, July 1997.
- 4 C. Cortes, Marius Kloft, and M. Mohri. Learning kernels using local rademacher complexity. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 2013. in press.
- 5 L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer svm for video concept detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- 6 L. Duan, D. Xu, I. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- 7 Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004.
- 8 Ali Farhadi and Mostafa Kamali Tabrizi. Learning to recognize activities from the wrong view point. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008.
- 9 Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, July 2006.
- 10 Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko. Efficient learning of domain-invariant image representations. In *International Conference on Learning Representations (ICLR)*, 2013.
- 11 Marius Kloft and Gilles Blanchard. On the convergence rate of ℓ_p -norm multiple kernel learning. *Journal of Machine Learning Research*, 13:2465–2502, Aug 2012.
- 12 Marius Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Lp-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, Mar 2011.
- 13 G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett, and M. I. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.
- 14 Fei-Fei Li, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.
- 15 Limin Li, Barbara Rakitsch, and Karsten M. Borgwardt. ccsvm: correcting support vector machines for confounding factors in biological data classification. *Bioinformatics [ISMB/ECCB]*, 27(13):342–348, 2011.
- 16 Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nat Meth*, 8(10):833–835, October 2011.
- 17 Jie Luo, Tatiana Tommasi, and Barbara Caputo. Multiclass transfer learning from unconstrained priors. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1863–1870, 2011.
- 18 Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.
- 19 Ariadna Quattoni, Michael Collins, and Trevor Darrell. Transfer learning for image classification with sparse prototype representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.
- 20 Erik Rodner and Joachim Denzler. Learning with few examples by transferring feature relevance. In *Proceedings of the 31st Annual Symposium of the German Association for Pattern Recognition (DAGM)*, pages 252–261, 2009.

- 21 Erik Rodner and Joachim Denzler. One-shot learning of object categories using dependent gaussian processes. In *Proceedings of the 32nd Annual Symposium of the German Association for Pattern Recognition (DAGM)*, pages 232–241, 2010.
- 22 Erik Rodner and Joachim Denzler. Learning with few examples for binary and multi-class classification using regularization of randomized trees. *Pattern Recognition Letters*, 32(2):244–251, 2011.
- 23 Ulrich Rückert and Marius Kloft. Transfer learning with adaptive regularizers. In *ECML/PKDD (3)*, pages 65–80, 2011.
- 24 Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision (ECCV)*, pages 213–226, 2010.
- 25 Gabriele Schweikert, Christian Widmer, Bernhard Schölkopf, and Gunnar Rätsch. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *Advances in Neural Information Processing Systems 21*, pages 1433–1440, 2009.
- 26 S. Sonnenburg, A. Zien, and G. Rätsch. Arts: Accurate recognition of transcription starts in human. *Bioinformatics*, 22(14):e472–e480, 2006.
- 27 Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.
- 28 Ingo Steinwart, Don R. Hush, and Clint Scovel. Learning from dependent observations. *J. Multivariate Analysis*, 100(1):175–194, 2009.
- 29 Antonio Torralba and Alyosha Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- 30 Antonio Torralba, Kevin P Murphy, and William T Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages II–762. IEEE, 2004.
- 31 C. Widmer, M. Kloft, and G. Rätsch. Multi-task learning for computational biology: Overview and outlook. In *B. Schoelkopf, Z. Luo, and V. Vovk, editors, Empirical Inference – Festschrift in Honor of Vladimir N. Vapnik*, 2013.

2 Table of Contents

Executive Summary

Erik Rodner, Trevor Darrell, Marius Kloft, Massimiliano Pontil, and Gunnar Rätsch 18

Overview of Talks

Transfer Learning using Marginal Distribution Information <i>Gilles Blanchard</i>	26
Non-i.i.d. Deep Learning <i>Trevor Darrell, Kate Saenko, Judy Hoffman</i>	27
Computer Vision to Support Decision Making in Ecology <i>Joachim Denzler</i>	31
Reproducing Kernel Hilbert Space Embeddings in Computational Biology <i>Philipp Drewe</i>	34
Bridging the Gap Between Synthetic and Real Data <i>Mario Fritz</i>	34
On the Need of Theory and Algorithms Correcting for Confounding Factors <i>Marius Kloft</i>	36
Transfer Learning in Computer Vision <i>Christoph H. Lampert</i>	38
Optimization for Machine Learning – Made Easy yet Efficient <i>Soeren Laue</i>	39
Transfer and Multi-Task Learning in Reinforcement Learning <i>Alessandro Lazaric</i>	40
Deep unsupervised domain adaptation by backpropagation <i>Victor Lempitsky</i>	42
Feature Learning in a Probit Model with Correlated Noise <i>Stephan Mandt</i>	44
A Resampling Method for Importance Weight Estimation <i>Shinichi Nakajima</i>	45
Not IID Data in Advertising <i>Francesco Orabona</i>	45
The Benefit of Multitask Representation Learning <i>Massimiliano Pontil</i>	46
Adaptive Lifelong Learning for Visual Recognition and Data Analysis <i>Erik Rodner</i>	46
Covariate Shift and Varying-Coefficient Models <i>Tobias Scheffer</i>	48
Kernel Hypothesis Tests on Dependent Data <i>Dino Sejdinovic</i>	50
Zero-shot learning via synthesized classifiers <i>Fei Sha</i>	51

A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems
with an Application to Learning
Ingo Steinwart 52

Sampling without replacement: direct approach vs. reduction to i.i.d.
Ilya Tolstikhin 52

Active Learning for Domain Adaptation
Ruth Urner 54

Working Groups, Presentations, and Panel Discussion 54

Participants 55

3 Overview of Talks

3.1 Transfer Learning using Marginal Distribution Information

Gilles Blanchard (University of Potsdam, DE)

License  Creative Commons BY 3.0 Unported license
© Gilles Blanchard

Consider a setting where a large number N of labeled training samples $S_i := (X_{ij}, Y_{ij})_{1 \leq j \leq n_i}$ ($i = 1, \dots, N$) on $\mathcal{X} \times \mathcal{Y}$ are available. The primary goal is not to find an adequate classification (or regression) function for each of these samples, but rather to find an appropriate prediction function $f : \mathcal{X} \times \mathcal{Y}$ for a *new, unlabeled* test sample $S^T := (X_j^T)_{1 \leq j \leq n^T}$. Such a situation occurs, for instance, for the *automatic gating* problem for flow cytometry data, a high-throughput measurement platform that is an important clinical tool for the diagnosis of many blood-related pathologies. The index i indicates a particular patient; for each patient a blood sample is taken, and measured by the device. This blood sample contains n_i individual cells – potentially several dozens of thousands – each of which is separately analyzed by the device, giving rise to a feature vector X_{ij} of attributes related to physical and chemical properties of the individual cell. The label Y_{ij} , input manually by an expert, gives the type of each cell (blood cell, white cell, etc.). The goal is to make this last labelling (or “gating”) step automatic, using the available labeled data. Note that in this case, for a new test patient zero label information is available, only the feature vectors of the cells present in the blood sample.

This problem belongs to the vast landscape of transfer learning. A classical approach to the problem (the covariate shift setting) assumes that the marginal distribution $P_X^{(i)}$ differs between samples, but that the conditional $P_{Y|X}$ stays the same. This is a very strong assumption that we want to avoid. We propose the following alternative Ansatz: there is a relationship common to all samples from the marginal $P_X^{(i)}$ to the conditional $P_{Y|X}^{(i)}$. Thus, we posit that there is some pattern making it possible to learn a mapping from marginal distributions to labels. We call this setting *marginal predictor learning*. In other words, we want to learn a mapping

$$f : \mathfrak{P}_{\mathcal{X}} \times \mathcal{X} \rightarrow \mathbb{R},$$

(where $\mathfrak{P}_{\mathcal{X}}$ denotes the set of marginal distributions on \mathcal{X}) which, for a new unlabeled sample with corresponding empirical marginal distribution \widehat{P}_X^T , will predict the label $f(\widehat{P}_X^T, x)$ for a specific feature vector x belonging to that sample.

We show that this setting is amenable to a reproducing kernel learning method. The gist of our approach is to combine recent developments about kernels on distributions (Christmann and Steinwart 2010, Sriperumbudur et al. 2010) with ideas of kernel multitask learning (Evgeniou and Pontil 2005). In a nutshell, the abstract “kernel task similarity matrix” present in kernel multitask learning is replaced by a task similarity matrix determined by the similarity between empirical marginal distributions, as measured by a distribution kernel. We show in particular that this approach is universally consistent under weak assumptions, is practically applicable and can outperform other approaches.

References

- 1 G. Blanchard, G. Lee, C. Scott. Generalizing from Several Related Classification Tasks to a New Unlabeled Sample. In *NIPS*, 2178-2186, 2011.
- 2 A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. In *NIPS*, 406-414, 2010.

- 3 T. Evgeniou and M. Pontil. Learning multiple tasks with kernel methods. In *JMLR* (6), 615–637, 2005.
- 4 B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. In *JMLR* (11), 1517–1561, 2010.

3.2 Non-i.i.d. Deep Learning

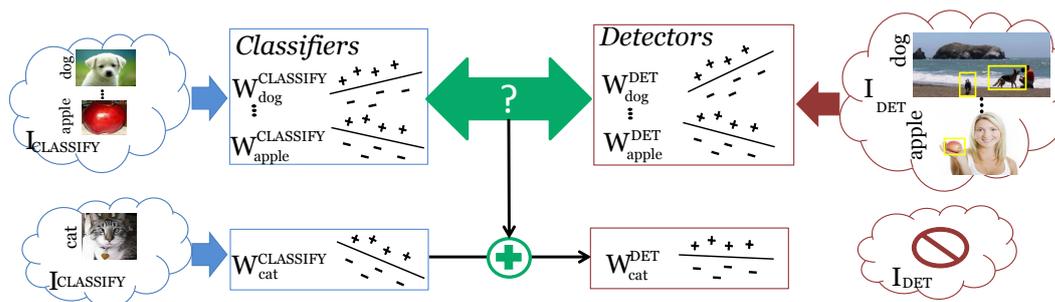
Trevor Darrell (UC Berkeley, US), Kate Saenko (UMass Lowell, US), Judy Hoffman (UC Berkeley, US)

License © Creative Commons BY 3.0 Unported license
© Trevor Darrell, Kate Saenko, Judy Hoffman

LSDA: Detection as Domain Adaptation

One of the fundamental challenges in training object detection systems is the need to collect a large amount of images with bounding box annotations. The introduction of detection challenge datasets, such as PASCAL VOC [9], have propelled progress by providing the research community a dataset with enough fully annotated images to train competitive models although only for 20 classes. Even though the more recent ImageNet detection challenge dataset [3] has extended the set of annotated images, it only contains data for 200 categories. As we look forward towards the goal of scaling our systems to human-level category detection, it becomes impractical to collect a large quantity of bounding box labels for tens or hundreds of thousands of categories.

We ask, is there something generic in the transformation from classification to detection that can be learned on a subset of categories and then transferred to other classifiers? We cast this task as a domain adaptation problem, considering the data used to train classifiers (images with category labels) as our source domain, and the data used to train detectors (images with bounding boxes and category labels) as our target domain. We then seek to find a general transformation from the source domain to the target domain, that can be applied to any future classifier to adapt it into a detector.

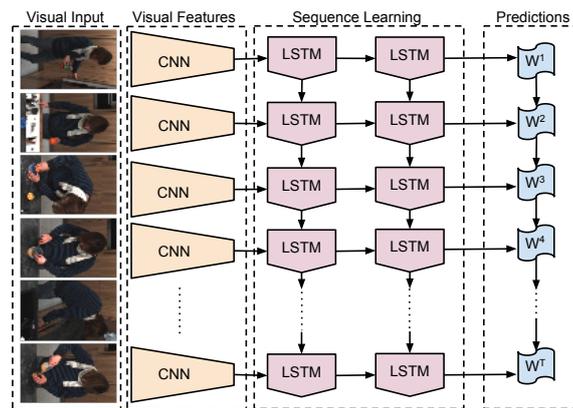


■ **Figure 1** The core idea is that we can learn detectors (weights) from labeled classification data (left), for a wide range of classes. For some of these classes (top) we also have detection labels (right), and can learn detectors. But what can we do about the classes with classification data but no detection data (bottom)? Can we learn something from the paired relationships for the classes for which we have both classifiers and detectors, and transfer that to the classifier at the bottom to make it into a detector?

We have already released a 7.6K visually grounded lexicon comprised of detectors adapted from ImageNet classifiers, available at <https://github.com/jhoffman/llda>. Our model is based

on a technique we call Large Scale Detection through Adaptation (LSDA), an algorithm that learns to transform an image classifier into an object detector [15]. To accomplish this goal, we use supervised convolutional neural networks (CNNs), which have recently been shown to perform well both for image classification [18] and object detection [10, 21]. We have recently extended this model to also solve a latent variable task to identify inlier visual regions, further improving learning from images of complex scenes [16].

In the future, we will extend this model beyond its present formulation based on WordNet to include similar concepts which can be learned from static imagery, including adjectives, and to incorporate motion representations for learning verbs. E.g., we hope to provide groundings similar to that in the Columbia “adjective noun pairs” dataset of [6], but integrated into the LSDA detector framework.



■ **Figure 2** We introduced *Long-term Recurrent Convolutional Networks* (LRCNs), a class of architectures leveraging the strengths of rapid progress in CNNs for visual recognition problem, and the growing desire to apply such models to time-varying inputs and outputs. This enables learning from images and videos with only weak labels in the form of tags or captions. LRCN processes the (possibly) variable-length visual input (left) with a CNN (middle-left), whose outputs are fed into a stack of recurrent sequence models (*LSTMs*, middle-right), which finally produce a variable-length prediction (right). Please see goo.gl/cZRM4U for example output sentences.

LRCN: Weak learning from images, videos, and captions

Image data collection for individual concepts may have reached a plateau in productivity, and we predict stronger models will result from models which leverage images and text in context, with only indirect labeling. Learning models from images or videos and associated captions or descriptive text is an especially appealing method for grounding elementary units in perceptual experience, as the system learns how to align image and textual content without explicit supervision.

Recognition and description of images and videos is a fundamental challenge of computer vision. Dramatic progress has been achieved by supervised convolutional models on image recognition tasks, and a number of extensions to process video have been recently proposed. Ideally, a video model should allow processing of variable length input sequences, and also provide for variable length outputs, including generation of full-length sentence descriptions that go beyond conventional one-versus-all prediction tasks. We have produced *long-term recurrent convolutional networks* (LRCNs), a novel architecture for visual recognition and

description which combines convolutional layers and long-range temporal recursion and is end-to-end trainable (see Figure 2).

We have instantiated our architecture for specific video activity recognition, image caption generation, and video description tasks. We have shown that long-term recurrent convolutional models are generally applicable to visual time-series modeling and that these models improve generation of descriptions from intermediate visual representations derived from conventional visual models. We instantiate our proposed architecture in three experimental settings. First, we show that directly connecting a visual convolutional model to deep LSTM networks, we are able to train video recognition models that capture complex temporal state dependencies. While existing labeled video activity datasets may not have actions or activities with extremely complex time dynamics, we nonetheless see improvements on the order of 4% on conventional benchmarks. and importantly enable direct end-to-end trainable image-to-sentence mappings. Strong results for machine translation tasks have recently been reported [22, 7]; such models are encoder/decoder pairs based on LSTM networks. Our multimodal architecture consists of a visual CNN to encode a deep state vector and an LSTM to decode the vector into a natural language string. This model can be trained end-to-end on large-scale image and text datasets, and even with modest training provides competitive generation results compared to existing methods.

To date, there has only been limited investigation of what has been learned in these models, and little systematic exploration of how such knowledge can be extracted and leveraged in related tasks. Anecdotal results suggest that the LCRN model does learn how to localize specific noun phrases and can learn to ground complex and/or idiosyncratic terms.

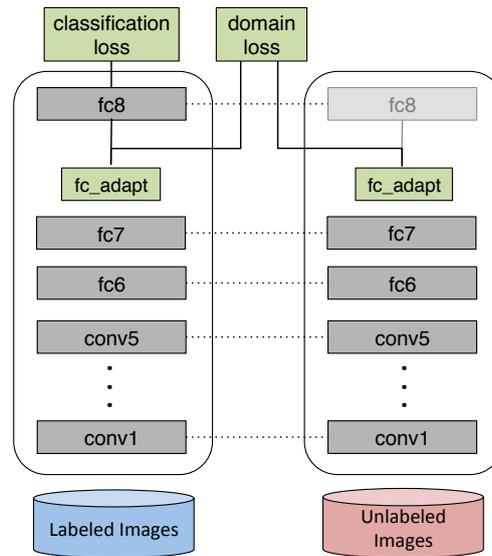
We propose to combine the variable input weak learning model with our large scale detection through adaptation approach to create models that not only produce captions and descriptions for novel videos/images, but are also able to localize the salient nouns and verbs. This will enable interactive applications and provide an intuitive medium through which to communicate with users.

Towards Deep Confusion

The methods proposed above presume a (possibly weakly labeled) supervised learning regime, with test and training data coming from the same domain. It is a widely recognized phenomenon that models trained in one environment, even with large data sources, suffer from degraded performance when deployed in a new or specialized environment. For example, a model trained on web search images may not perform very well for recognition on a robot mounted camera in a warehouse or office environment. In order for our large scale models to be widely applicable, we will develop algorithms that quickly adapt to new scenarios without the expensive overhead of collecting new labeled data and retraining a model from scratch.

Dataset bias is a well known and theoretically understood problem with traditional supervised approaches to image recognition [23]. A number of recent theoretical and empirical results have shown that supervised methods' test error increases in proportion to the difference between the test and training input distribution [2, 4, 20, 23]. In the last few years, several methods for visual domain adaptation have been suggested to overcome this issue [8, 24, 1, 20, 19, 17, 12, 11, 13, 14], but were limited to shallow models. The traditional approach to adapting deep models has been fine-tuning; see [10] for a recent example.

We propose a new CNN architecture, outlined in Figure 3, which uses an adaptation layer along with a domain confusion loss based on maximum mean discrepancy (MMD) [5] to automatically learn a representation jointly trained to optimize for classification and domain invariance. Our domain confusion metric can be used both to select the dimension



■ **Figure 3** Our architecture optimizes a deep CNN for both classification performance and domain invariance. The model can be trained for *supervised* adaptation, when there is a small number of target labels available, or *unsupervised* adaptation, when no target labels are available. We introduce domain invariance through *domain confusion* guided selection of the depth and width of the adaptation layer, as well as an additional loss term during fine-tuning that directly minimizes the distance between source and target representations.

of the adaptation layers, choose an effective placement for a new adaptation layer within a pre-trained CNN architecture, and fine-tune the representation. Our architecture can be used to solve both *supervised adaptation*, when a small amount of target labeled data is available, and *unsupervised adaptation*, when no labeled target training data is available.

References

- 1 Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In *Proc. ICCV*, 2011.
- 2 Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Proc. NIPS*, 2007.
- 3 A. Berg, J. Deng, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. 2012.
- 4 John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Proc. NIPS*, 2007.
- 5 Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *Bioinformatics*, 2006.
- 6 Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM, 2013.
- 7 Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.
- 8 H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, 2007.

- 9 M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.
- 10 R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *In Proc. CVPR*, 2014.
- 11 B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proc. CVPR*, 2012.
- 12 R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Proc. ICCV*, 2011.
- 13 J. Hoffman, B. Kulis, T. Darrell, and K. Saenko. Discovering latent domains for multisource domain adaptation. In *Proc. ECCV*, 2012.
- 14 J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain-invariant image representations. In *Proc. ICLR*, 2013.
- 15 Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pages 3536–3544, 2014.
- 16 Judy Hoffman, Deepak Pathak, Trevor Darrell, and Kate Saenko. Detector discovery in the wild: Joint multiple instance and representation learning, 2015.
- 17 A. Khosla, T. Zhou, T. Malisiewicz, A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *Proc. ECCV*, 2012.
- 18 A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.
- 19 B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proc. CVPR*, 2011.
- 20 K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proc. ECCV*, 2010.
- 21 P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.
- 22 Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.
- 23 A. Torralba and A. Efros. Unbiased look at dataset bias. In *Proc. CVPR*, 2011.
- 24 J. Yang, R. Yan, and A. Hauptmann. Adapting SVM classifiers to data with shifted distributions. In *ICDM Workshops*, 2007.

3.3 Computer Vision to Support Decision Making in Ecology

Joachim Denzler (Friedrich Schiller University Jena, DE)

License © Creative Commons BY 3.0 Unported license
© Joachim Denzler

Ecology is the study of life and its interaction with the physical environment. Scientists are interested in quantifying relations between atmospheric, oceanic, and terrestrial processes. For a long time, analysis has been done locally both with respect to the region of investigation as well as with respect to the field in which phenomena are studied. Due to the possibilities to record data all over the world, the increase in resolution, the quality of recordings from satellites, distributions of data sets over the world wide web, and computing in the cloud new opportunities arise. Such heterogenous and globally collected data may make it possible to answer questions that are of fundamental importance for the future of our planet.

In this research domain computer vision can play an important role in the future. Today, most work by researchers in ecology is done by analyzing data manually. For example, the number of butterflies in a certain region is determined by visual inspection of traps installed in the environment.

Over the last years, computer vision research already tackled problems that are of high relevance for ecology as well. One example is the automatic analysis of remote sensing data. A second example is the identification of animals from images and videos. Birds, dogs, mushrooms, flowers build databases for object recognition benchmarks, since those objects not just offer very challenging problems but also call for new methods, that lead to the area of fine-grained recognition. Works directly related to ecology are for example, the classification of insects [1], or computer vision methods for coral reef assessment [2].

One hypothesis of our research is that computer vision methods can only be accepted and successful in ecology, if we are able to exploit all knowledge (labeled data from similar domains, common feature representations, etc.) already available, to incrementally improve performance, and to keep the human in the loop, for example, to check of correct automatic decision. This allows to build automatic systems with minimal user efforts – a preliminary, if researchers from other disciplines shall accept modern techniques from computer vision for their research. Domain adaptation and transfer learning will play one key role to success.

When working together with people from ecology and biodiversity research, specific problems arise that must be solved from the computer vision and machine learning perspective:

1. can we configure initial classifiers for ecology applications using already existing data bases or images from the internet?
2. can we adapt existing classifiers using a minimal set of training data from a specific application scenario, to reduce the effort by researchers from ecology?
3. can the process of domain adaptation be supported by the human in the loop, for example, to embed it into a life-long learning scenario?
4. can we exploit data from additional modalities besides visual data to support transfer learning in the visual domain?
5. are there common principles in transfer learning that can also be applied to analyse dynamic processes, for example, the interactions between animals – with special focus on behaviour changing over time
6. can we benefit from the huge amount of data that will be collected in the future, and are existing methods from machine learning already capable to deal with streams of input data for model update

The Computer Vision Group Jena aims at life-long learning scenarios, including large scale visual learning and recognition [3], active learning [4, 5], novelty detection [6, 7, 8], incremental learning [9], and fine-grained recognition [10, 11]. For dynamic scene analysis, computer vision in sensor networks has been one goal during the past years as well, with the focus on supervised and unsupervised activity recognition [12, 13]. Applications so far came from biology (unsupervised mytosis detection [14]) and medicine (classification of facial paralysis [15]). In the later case we investigated domain adaptation for active appearance models [16].

The Computer Vision Group, headed by Joachim Denzler, consists of two senior researchers, Erik Rodner and Wolfgang Ortmann, and currently 12 PhD students. Joachim is also faculty member of the Abbe-School of Photonics and the International Max-Planck-Research School for Global Biochemical Cycles. He is co-founder of the Michael Stifel Center for Data-Driven and Simulation Science Jena.

References

- 1 N. Larios, B. Soran: Haar random forest features and svm spatial matching kernel for stonefly species identification. In: International Conference on Pattern Recognition. (2010)
- 2 Beijbom, O., Edmunds, P.J., Kline, D.I., Mitchell, B.G., Kriegman, D.: Automated annotation of coral reef survey images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, Rhode Island (2012)
- 3 Fröhlich, B., Rodner, E., Kemmler, M., Denzler, J.: Large-scale gaussian process multi-class classification for semantic segmentation and facade recognition. *Machine Vision and Applications* **24**(5) (2013) 1043–1053
- 4 Käding, C., Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Active learning and discovery of object categories in the presence of unnameable instances. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
- 5 Freytag, A., Rodner, E., Denzler, J.: Selecting influential examples: Active learning with expected model output changes. In: European Conference on Computer Vision (ECCV). Volume 8692. (2014) 562–577
- 6 Bodesheim, P., Rodner, E., Freytag, A., Denzler, J.: Divergence-based one-class classification using gaussian processes. In: British Machine Vision Conference (BMVC). (2012) 50.1–50.11
- 7 Bodesheim, P., Freytag, A., Rodner, E., Kemmler, M., Denzler, J.: Kernel null space methods for novelty detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013)
- 8 Bodesheim, P., Freytag, A., Rodner, E., Denzler, J.: Local novelty detection in multi-class recognition problems. In: IEEE Winter Conference on Applications of Computer Vision (WACV). (2015) 813–820
- 9 Lütz, A., Rodner, E., Denzler, J.: I want to know more: Efficient multi-class incremental learning using gaussian processes. *Pattern Recognition and Image Analysis* **23**(3) (2013) 402–407
- 10 Simon, M., Rodner, E., Denzler, J.: Fine-grained classification of identity document types with only one example. In: Machine Vision Applications (MVA). (2015)
- 11 Göring, C., Rodner, E., Freytag, A., Denzler, J.: Nonparametric part transfer for fine-grained recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) 2489–2496
- 12 Krishna, M.V., Denzler, J.: A combination of generative and discriminative models for fast unsupervised activity recognition from traffic scene videos. In: IEEE Winter Conference on Applications of Computer Vision (WACV). (2014) 640–645
- 13 Körner, M., Denzler, J.: Temporal self-similarity for appearance-based action recognition in multi-view setups. In: *Computer Analysis of Images and Patterns*. Volume 8047. (2013) 163–171
- 14 Krishna, M.V., Denzler, J.: A hierarchical bayesian approach for unsupervised cell phenotype clustering. In: German Conference on Pattern Recognition (GCPR). (2014) 69–80
- 15 Haase, D., Kemmler, M., Guntinas-Lichius, O., Denzler, J.: Efficient measuring of facial action unit activation intensities using active appearance models. In: IAPR International Conference on Machine Vision Applications (MVA). (2013) 141–144
- 16 Haase, D., Rodner, E., Denzler, J.: Instance-weighted transfer learning of active appearance models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) 1426–1433

3.4 Reproducing Kernel Hilbert Space Embeddings in Computational Biology

Philipp Drewe (Max-Delbrück-Centrum, DE)

License © Creative Commons BY 3.0 Unported license
© Philipp Drewe

Joint work of Drewe, Philipp; Stegle, Oliver; Hartmann, Lisa; Kahles, André; Bohnert, Regina; Wachter, Andreas; Borgwardt, Karsten; Rätsch, Gunnar

Main reference P. Drewe, O. Stegle, L. Hartmann, A. Kahles, R. Bohnert, A. Wachter, K. Borgwardt, G. Rätsch, “Accurate detection of differential RNA processing,” *Nucleic Acids Research*, 41(10):5189–5198, 2013.

URL <http://dx.doi.org/10.1093/nar/gkt211>

A fundamental problem in computational biology is identifying genes in a cell that are processed differently upon perturbation of the cell. However, this is challenging as the processing of the genes cannot be directly measured, but has to be inferred from a set of incomplete observations (reads) of the genes. These reads are high-dimensional, structured and typically non-iid distributed. Therefore, classical statistical test, such as the Kolmogorov-Smirnov test, cannot be applied in this setting. In this work, we show that Reproducing Kernel Hilbert Space (RKHS) embeddings allow a suitable representation of read-data. Furthermore, we present RKHS-embedding-based approaches to test for homogeneity of two sets of observations, in order to accurately identify genes whose processing has changed.

3.5 Bridging the Gap Between Synthetic and Real Data

Mario Fritz (MPI für Informatik – Saarbrücken, DE)

License © Creative Commons BY 3.0 Unported license
© Mario Fritz

There is a long tradition of using generative models in combination with discriminative classifiers [5, 6, 7]. Equally the recently successful deep learning technique [3] use jittering techniques [1, 2] that imply sampling from an underlying distribution. Although in both cases the the model is postulated and all parameters are in our control, we rarely achieve an accurate representation of the true underlying distribution. Yet, these techniques have shown improved performance as learning is guided by prior knowledge encoded in such generative models.

Learning and Prediction from Rendered/Synthesized Data

Many applications greatly benefit by means of synthesizing additional training data. For visual recognition this often involves a rendering process for creating new images. The employed model represents prior knowledge about the target domain. In this section, several examples are listed where we have directly used the rendered data – assuming that the domain mismatch between real and virtual examples is negligible.

Detection by Rendering. In early work, we have captured a light-field of an object and rendered new views of the object on demand in order to evaluate the posterior in a particle filter tracking framework [8].

New View Synthesis. Human generalize easily from a single view of an object to novel view-points. Today’s computer vision algorithms are mostly learning and example based and therefore have to be shown variations across style and viewpoints in order to succeed. We

have presented an approach that uses a 3D model to guide novel view synthesis, that is able to fill in disocclusion areas truthfully [9]. The object models trained on such augmented data show a greatly improved view point generalization.

Differentiable Vision Pipeline. Most recently, we have established a fully differentiable vision pipeline [10] that builds on top of an approximately differentiable renderer [4] and a differentiated HOG image representation. This allows us to estimate object poses by exploiting the prescribed image synthesis procedure in the gradient computation.

Adaptation to Rendered/Synthesized Data

Although significant progress has been achieved by solely relying on realistic rendering and synthesis, quite often the domain shift between the virtual and the real world introduces a distribution mismatch that should be treated separately.

Visual Domain Adaptation via Metric Learning. We have proposed to reduce the effects of domain shifts by a metric learning formulation [11]. Hereby we have improved recognition across different data sources such a webcam, dslr or data from the web.

Recognition from Virtual Examples. We have employed the concept of metric learning for domain adaptation to the problem of visual material recognition [12]. The approach helps to bridge the gap between rendered and real data.

Prediction under changing prior distribution. Most recently, we have shown how to perform gaze estimation in the wild [13]. Considering the change in the prior distribution of head pose and eye fixation distribution has been critical when training across datasets.

Unsupervised Adaptation

Future challenges include scenarios where no training data for adaptation is available. Less work has been performed in this direction. We have proposed to adapt to new conditions in a road segmentation task by assuming a stationary, structured prior over the label space, which allows us to successfully adapt a semantic labeler to unseen weather conditions [14]. Beyond the traditional recognition scenarios, we have also attempted to bring the required adaptivity to learning settings. E.g. we have adapted active learning strategies via reinforcement learning to different training distributions [15]. We hypothesize that non-parametric learning techniques for visual recognition and grouping [16] can be well suited to transfer structural relations across domains, while being less affected by changes in individual appearances.

References

- 1 P. Simard, B. Victorri, Y. LeCun, J. Denker. Tangent prop-a formalism for specifying selected invariances in an adaptive network. In *Advances in neural information processing systems (NIPS)*, 1992
- 2 D. Decoste, B. Schölkopf. Training invariant support vector machines. In *Journal of Machine Learning*, 2002
- 3 A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.
- 4 M. Loper, M. Black. Opendr: An approximate differentiable renderer. In *European Conference on Computer Vision (ECCV)*, 2014
- 5 T. Jaakkola, D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems (NIPS)*, 1999

- 6 M. Fritz, B. Leibe, B. Caputo, B. Schiele. Integrating representative and discriminant models for object category detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2005
- 7 A. Holub, M. Welling, P. Perona. Combining generative models and fisher kernels for object recognition. In *IEEE International Conference on Computer Vision (CVPR)*, 2005
- 8 M. Zobel, M. Fritz, and I. Scholz. Object tracking and pose estimation using light-field object models. In *Vision, Modeling, and Visualization Conference (VMV)*, 2002.
- 9 K. Rematas, T. Ritschel, M. Fritz, and T. Tuytelaars. Image-based synthesis and re-synthesis of viewpoints guided by 3d models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- 10 W.-C. Chiu and M. Fritz. See the difference: Direct pre-image reconstruction and pose estimation by differentiating hog. *arXiv:1505.00663 [cs.CV]*, 2015.
- 11 K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision (ECCV)*, 2010.
- 12 W. Li and M. Fritz. Recognizing materials from virtual examples. In *European Conference on Computer Vision (ECCV)*, 2012.
- 13 X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- 14 E. Levinkov and M. Fritz. Sequential bayesian model update under structured scene prior for semantic road scenes labeling. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.
- 15 S. Ebert, M. Fritz, B. Schiele. Ralf: A reinforced active learning formulation for object class recognition In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- 16 W.-C. Chiu, M. Fritz. Multi-class video co-segmentation with a generative multi-video model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013

3.6 On the Need of Theory and Algorithms Correcting for Confounding Factors

Marius Kloft (HU Berlin, DE)

License  Creative Commons BY 3.0 Unported license
© Marius Kloft

A classic assumption in machine learning states that the data is independently realized from an unknown distribution. This assumption greatly simplifies theory [1] and algorithms [2]. However, it is common in several applications that the data exhibit dependencies and inherent correlations between observations. Clearly, this occurs especially for time series, for instance, in network security (e.g., HTTP requests) and computer vision (video streams). Under the assumption of time-structured dependencies, several algorithms and theory have been proposed [3]. But few theory and algorithms have been developed for complex dependencies, in particular for confounding ones.

For instance in statistical genetics, it is one of the central challenges to detect – among ten thousands of genes – the ones that are strong predictors of complex diseases or other binary outcomes [4, 5], as it is a first step in identifying regulatory components controlling heritability. However, for various diseases such as type 2 diabetes [6], these sparse signals are yet largely undetected, which is why these missing associations have been entitled the *The Dark Matter of Genomic Associations* [7]. Central problems include that these signals are

often very weak, and the found signals can be spurious due to confounding. Confounding can stem from varying experimental conditions and demographics such as age, ethnicity, gender [8], and – crucially – population structure, which is due to the relatedness between the samples [9, 8, 10]. Ignoring such confounders can often lead to spurious false positive findings that cannot be replicated on independent data [11]. Correcting for such confounding dependencies is considered one of the greatest challenges in statistical genetics [12]. Another example is content- and anomaly-based network intrusion detection and malware detection, where attacks are recorded within sandboxes [13]. Thus attributes that are specific to sandboxes help in discriminating attacks from benign data so that these attributes may be falsely promoted by the learning algorithm.

In the present Dagstuhl workshop, we found that there is a lack of research in the above respect. Which is why we advocate to develop theory and algorithms learning and estimation in the presence of confounding, the basic aim of which would be to understand and create statistical machine learning from confounded data. In particular, the following open problems arose at the workshop:

- How can we quantify “confoundedness” in learning settings?
- Can we develop theory similar to uniform convergence kind of analyses [1] under the assumption of confounders? And in order for this to work which assumptions do we need to state?
- How to design effective learning algorithms in presence of confounding and dependent labels?
- How to address feature selection under confounders?
- How to automatically learn the confounders?

Addressing the above stated open questions will subject to interesting future work. A good starting point to this end will be previous theoretical analyses regarding time series [3, 14] and probabilistic models such as the probit regression model [15, 16], and its extensions to GP classification [17, 18] and generalized linear mixed models [19].

Acknowledgments. MK acknowledges support by the German Research Foundation (DFG) under KL 2698/2-1.

References

- 1 V. N. Vapnik and A. Y. Chervonenkis, “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities,” *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264–280, 1971.
- 2 C. Cortes and V. Vapnik, “Support vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995.
- 3 M. Mohri and A. Rostamizadeh, “Rademacher complexity bounds for non-i.i.d. processes,” in *NIPS*, pp. 1097–1104, 2008.
- 4 T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, *et al.*, “Finding the missing heritability of complex diseases,” *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.
- 5 S. Vattikuti, J. J. Lee, C. C. Chang, S. D. Hsu, and C. C. Chow, “Applying compressed sensing to genome-wide association studies,” *GigaScience*, vol. 3, no. 1, p. 10, 2014.
- 6 N. Craddock, M. E. Hurles, N. Cardin, *et al.*, “Genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls,” *Nature*, vol. 464, no. 7289, pp. 713–720, 2010.
- 7 T. N. H. G. R. Institute, “Proceedings of the workshop on the dark matter of genomic associations with complex diseases: Explaining the unexplained heritability from genome-wide association studies,” 2009.

- 8 L. Li, B. Rakitsch, and K. M. Borgwardt, “ccsvm: correcting support vector machines for confounding factors in biological data classification,” *Bioinformatics*, vol. 27, no. 13, pp. 342–348, 2011.
- 9 C. Lippert, J. Listgarten, Y. Liu, C. Kadie, R. Davidson, and D. Heckerman, “Fast linear mixed models for genome-wide association studies,” *Nature Methods*, vol. 8, pp. 833–835, October 2011.
- 10 N. Fusi, O. Stegle, and N. D. Lawrence, “Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical studies,” *PLoS comp. bio.*, vol. 8, no. 1, 2012.
- 11 P. Kraft, E. Zeggini, and J. P. Ioannidis, “Replication in genome-wide association studies,” *Statistical Science: A review journal of the Institute of Mathematical Statistics*, vol. 24, no. 4, p. 561, 2009.
- 12 B. J. Vilhjálmsson and M. Nordborg, “The nature of confounding in genome-wide association studies,” *Nature Reviews Genetics*, vol. 14, no. 1, pp. 1–2, 2013.
- 13 D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, K. Rieck, and C. Siemens, “Drebin: Effective and explainable detection of android malware in your pocket,” in *Proc. of NDSS*, 2014.
- 14 I. Steinwart, D. R. Hush, and C. Scovel, “Learning from dependent observations,” *J. Multivariate Analysis*, vol. 100, no. 1, pp. 175–194, 2009.
- 15 C. I. Bliss, “The method of probits,” *Science*, vol. 79, no. 2037, pp. 38–39, 1934.
- 16 L. Fahrmeir, T. Kneib, S. Lang, and B. Marx, *Regression*. Springer, 2013.
- 17 C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.
- 18 J. P. Cunningham, P. Hennig, and S. Lacoste-Julien, “Gaussian probabilities and expectation propagation,” *arXiv preprint arXiv:1111.6832*, 2011.
- 19 N. E. Breslow and D. G. Clayton, “Approximate inference in generalized linear mixed models,” *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 9–25, 1993.

3.7 Transfer Learning in Computer Vision

Christoph H. Lampert (IST Austria – Klosterneuburg, AT)

License © Creative Commons BY 3.0 Unported license

© Christoph H. Lampert

Joint work of Lampert, Christoph H.; Pentina, Anastatia; Sharmanska, Viktoriia

Main reference A. Pentina, C. H. Lampert, “A PAC-Bayesian Bound for Lifelong Learning,” in *Proc. of the 31th Int’l Conf. on Machine Learning (ICML’14)*, pp. 991–999, JMLR.org, 2014.

URL <http://jmlr.org/proceedings/papers/v32/pentina14.html>

URL <http://pub.ist.ac.at/~chl/erc>

Computer Vision offers a wide range of problems where transfer learning techniques, such as domain adaptation and multi-task learning, can be applied. Several such techniques have proven useful in practice, but a solid theoretical understanding of when and how transfer learning offer benefits for computer vision tasks is still lacking. In my research group at IST Austria, we are particularly interested in the problem of lifelong learning. A lifelong learner continuously and autonomously learns from a stream of data, potentially for years or decades [1, 2]. During this time the learner should build an ever-improving base of generic information, and use this as background knowledge and context for solving different tasks. Using PAC-Bayesian learning theory, we have developed theoretic foundations that allow us to study different lifelong learning situations [3]. The generalization bounds that we obtain consist only of computable quantities and can therefore be used to analyze existing

lifelong learning algorithms and derive new ones. Similar techniques also allow the analysis of algorithms for sequential multi-task learning [4].

Acknowledgments. The described work was funded by the European Research Council under the European Unions Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 308036.

References

- 1 S. Thrun and T. M. Mitchell. Lifelong robot learning. *Robotics and autonomous systems*, 15(1):25–46, 1995.
- 2 J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.
- 3 A. Pentina and C. H. Lampert. A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning (ICML)*, 2014.
- 4 A. Pentina, V. Sharmanska and C. H. Lampert. Curriculum Learning of Multiple Tasks. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

3.8 Optimization for Machine Learning – Made Easy yet Efficient

Soeren Laue (Friedrich Schiller University Jena, DE)

License © Creative Commons BY 3.0 Unported license
© Soeren Laue

Many machine learning problems are cast as continuous optimization problems. A non-exhaustive list of such problems includes support vector machines [2], elastic nets [8], dimension reduction [1], and sparse PCA [9]. Moreover, for a given machine learning problem there is typically not only a single formulation as an optimization problem but different formulations that, for example, take previous knowledge or constraints into account. In the case of support vector machines the original formulation uses an ℓ_2 -regularization term combined with the hinge loss. Different variants include the use of different loss functions, e.g., an ℓ_2 -loss term for adapting to Gaussian noise, ℓ_1 -regularization to obtain sparse predictors [7], or a combination of ℓ_1 - and ℓ_2 -regularization. Adding to this already large variety is the use of kernels in many of the problem formulations. However, up to this day, efficient solutions to any of these formulations still require the implementation of specialized, and highly-tuned solvers, not only in the case of support vector machines but for almost any machine learning problem that has been formulated as an optimization problem. This of course poses a problem when dealing with data sets whose size is well beyond the reach of easy to use modeling languages combined with a generic solver.

We present a novel approach to mitigate this problem by tightly coupling the modeling language and the generic solver. This results in code that is a few orders of magnitude more efficient than state-of-the-art modeling language/generic solver combinations like CVX/Gurobi [3, 4, 5] and CVX/Mosek [3, 4, 6]. The tight coupling is achieved by a generative programming approach that generates an individual solver for each problem as an instance of a generic solver. The generic optimizer is able to solve almost any continuous optimization problem with constraints over \mathbb{R}^n that has been proposed for machine learning tasks. It combines the ease of use of commonly used modeling languages with the efficiency of highly-tuned, specialized state-of-the-art solvers for the individual machine learning problems. In the end, the automatically generated solver can be either deployed as a callable library or as a stand-alone solver.

References

- 1 Christopher J. C. Burges. Dimension reduction: A guided tour. *Foundations and Trends in Machine Learning*, 2(4), 2010.
- 2 Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- 3 CVX Research, Inc. CVX: Matlab software for disciplined convex programming, version 2.0. <http://cvxr.com/cvx>, August 2012.
- 4 Michael Grant and Stephen Boyd. Graph Implementations for Nonsmooth Convex Programs. In *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag, 2008.
- 5 Gurobi Optimization, Inc. Gurobi Optimizer Reference Manual, 2013.
- 6 MOSEK ApS. The MOSEK Optimization Software, 2013.
- 7 Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, pages 267–288, 1996.
- 8 Hui Zou and Trevor Hastie. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, pages 301–320, 2005.
- 9 Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

3.9 Transfer and Multi-Task Learning in Reinforcement Learning

Alessandro Lazaric (INRIA Lille, FR)

License  Creative Commons BY 3.0 Unported license
© Alessandro Lazaric

The Context

Reinforcement learning's (RL) [5, 1] challenging objective is to develop autonomous agents able to learn how to act optimally in an unknown and uncertain environment by trial-and-error and with limited level of supervision (i.e., a reinforcement signal). RL is mostly applied in domains where a precise formalization of the environment and/or the efficient computation of the optimal control policy is particularly difficult (e.g., robotics, human-computer interaction, recommendation systems). An RL problem is formalized as a Markov decision process (MDP) \mathcal{M} characterized by a state space \mathcal{X} , an action space \mathcal{A} , a (stochastic) dynamics $p : \mathcal{X} \times \mathcal{A} \rightarrow \Delta(\mathcal{X})$ that determines the transition from states to states depending on the action, a reward function $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \rightarrow \mathbb{R}$ that determines the value of a transition x, a, x' . An MDP defines a control **task**. The solution to an MDP/task is an optimal policy $\pi^* : \mathcal{X} \rightarrow \mathcal{A}$ that prescribes the actions to take in each state to maximize the (discounted) sum of rewards measured by the optimal value function $V^* = \max_{\pi} \mathbb{E}[\sum_t \gamma^t r_t]$ with $\gamma \in (0, 1)$ and $r_t = r(x_t, \pi(x_t), x_{t+1})$. Two of the most difficult challenges in RL are:

1. How to explore the unknown environment so as to maximize the cumulative reward. This requires solving the **exploration-exploitation** problem, well formalized and studied at its core by the multi-armed bandit framework [2].
2. How to effectively represent the policy and/or the value function. This requires defining an **approximation space** which is well-suited for the specific MDP at hand.

Both previous aspects may greatly benefit from techniques able to define suitable exploration strategies and approximation spaces from past experience or joint experience from other tasks (e.g., designing an intelligent tutoring system for a student and reuse the teaching strategy to

other students). The **objective** of my research is to study the problems of transfer learning, multi-task learning, and domain adaptation in the RL (and related) field.

The Past

Unlike in supervised learning, transfer learning faces challenges which are specific to field of RL:

- many different things can be transferred (e.g., the MDP parameters, policies, value functions, samples, features),
- the definition of “unsupervised” samples is not clear and thus, domain adaptation methods exploiting target unsupervised samples cannot be easily applied,
- samples are often non-i.i.d. because they are obtained from policies
- tasks may be similar in terms of policies but neither MDPs nor value functions or viceversa.

For this reason, borrowing techniques from “supervised” transfer/multi-task learning is not always trivial or even possible. Early research focused on studying transfer of different kind of solutions from a source to a target task¹. Later, more sophisticated transfer/multi-task scenarios and algorithms have been developed (e.g., using hierarchical Bayesian solutions to learn “priors” from multiple tasks) to improve the accuracy of the approximation of optimal policies/value functions. The results obtained in the past show a significant sample complexity reduction and an improvement in asymptotic accuracy when transfer/multi-task is applied.

The Future

My main interest in the short-term is to study the problem of how transfer/multi-task learning can actually improve exploration-exploitation strategies in multi-armed bandit (MAB) and RL. While the problem of approximation is common in supervised learning as well, the active collection of information is very much specific to RL and MAB.

So far, I have investigated a sequential transfer scenario and investigated two approaches in the linear MAB framework: (i) transfer of samples (*under review*), (ii) use of transferred samples to identify the set of possible MAB problems and speed-up the problem identification phase [3]. In both cases, we proved that the cumulative reward (i.e., reduce the regret) of exploration-exploitation strategies in MAB can be actually improved and that negative transfer can be avoid. Nonetheless, a number of very important questions remain unanswered:

- Is it possible to incrementally and efficiently estimate the potential bias due to transfer from different tasks? Under which assumptions? *In specific cases, this can be done in supervised learning.*
- What is the measure of similarity between two MDPs that determines the difference in performance of an exploration-exploitation strategy when applied to the two MDPs?
- Is it worth it to explore more in earlier tasks to “unveil” the generative process of the sequence of tasks and exploit it to enhance the transfer? In which scenarios?
- MDPs with different state-action spaces may still be very much similar. Is it possible to map different MDP to an “underlying” common MDP structure in which similar exploration-exploitation solutions can be identified and transferred?

As motivating fields of application, I will focus on *intelligent tutoring systems, recommendation systems, and computer games*.

¹ See [6, 4] for a survey.

References

- 1 Bertsekas, D. and Tsitsiklis, J. *Neuro-Dynamic Programming*. Athena Scientific, 1996.
- 2 Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.
- 3 Gheshlaghi-Azar, M., Lazaric, A., and Brunskill, E. Sequential transfer in multi-arm bandit with finite set of models. In *Proceedings of the Twenty-Seventh Annual Conference on Neural Information Processing Systems (NIPS'13)*, 2013.
- 4 Lazaric, A. Transfer in reinforcement learning: a framework and a survey. In Wiering, M. and van Otterlo, M. (eds.), *Reinforcement Learning: State of the Art*. Springer, 2011.
- 5 Sutton, R. and Barto, A. *Reinforcement Learning, An introduction*. Bradford Book. The MIT Press, 1998.
- 6 Taylor, M. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10:1633–1685, 2000.

3.10 Deep unsupervised domain adaptation by backpropagation

Victor Lempitsky (Skoltech – Skolkovo, RU)

License © Creative Commons BY 3.0 Unported license
© Victor Lempitsky

Joint work of Ganin, Yaroslav; Lempitsky, Victor

Main reference Y. Ganin, V.S. Lempitsky, “Unsupervised Domain Adaptation by Backpropagation,” in Proc. of the 32nd Int’l Conf. on Machine Learning (ICML’15), pp. 1180–1189, JMLR.org, 2015.

URL <http://jmlr.org/proceedings/papers/v37/ganin15.html>

The method

Consider the problem of learning a deep feedforward classifier in the presence of domain shift. Assume that a large number of labeled source examples and a large number of unlabeled target examples are present (e.g. train on synthetic images, test on real one). Our approach [1] to this unsupervised domain adaptation problem is to combine deep learning and domain adaptation into a single optimization process driven by simple backpropagation updates. The goal of the optimization is to obtain a deep model that has domain-invariant feature representations in its higher layers, while providing good predictions on the source data.

Let \mathbf{x} be the input sample and y be the output of a network. Consider feature representation f that emerge after a certain layer L in the middle of the network. Let $\mathbf{f} = G_f(\mathbf{x}; \theta_f)$, $y = G_y(\mathbf{x}; \theta_y)$, where G_f and G_y are parts of the network before and after the layer L , while θ_f and θ_y are their parameters. Our goal is then to train a deep model where the features f are domain-invariant, i.e. have similar distribution in the source and the target domains. We denote these distributions as $S(\mathbf{f})$ and $T(\mathbf{f})$. While trying to match these distributions, one still needs to minimize the loss of the label prediction $y = G_y(G_f(\mathbf{x}; \theta_f); \theta_y)$ for source-domain data.

To measure the (dis)similarity of distributions $S(\mathbf{f})$ and $T(\mathbf{f})$, we augment our deep model with a domain classifier $d = G_d(\mathbf{f}; \theta_d)$. Given a feature vector \mathbf{f} this multi-layer classifier tries to predict whether it corresponds to the source or to the target example (i.e. whether it comes from $S(\mathbf{f})$ or $T(\mathbf{f})$). The lower is the loss of this classifier, the larger is the gap between $S(\mathbf{f})$ and $T(\mathbf{f})$. In the ideal case ($S(\mathbf{f})$ is the same as $T(\mathbf{f})$) this classifier would perform no better than chance and have a high loss. The resulting three-part network has a fork shape (forward pass through the network works as: $\mathbf{x} \rightarrow \mathbf{f}$, $\mathbf{f} \rightarrow y$, $\mathbf{f} \rightarrow d$).

The learning process trains all three parts of the network simultaneously using backpropagation. The training incorporates both labeled source examples and unlabeled target

examples. The parameters θ_y and θ_d are optimized by an SGD, with each update minimizing the losses of the respective classifiers G_y (that only looks at labeled source data) and G_d (that looks both at source and target data). The updates of the parameters θ_f of the feature mapping are driven by the minimization of the loss of the label predictor G_y and the *maximization* of the loss of the domain classifier G_d (as we want features to be predictive of y and domain-invariant).

We can achieve this behavior within standard deep learning packages based on SGD using a simple trick. We reverse (multiply by a negative constant) the gradient that comes out of the domain classifier G_d during backpropagation and pass it further back into the feature extractor. This can be implemented as a simple *gradient reversal* layer. When this layer is inserted between the feature extractor G_f and the domain classifier G_d , SGD moves the parameters θ_f against the direction suggested by the minimization of the domain classifier's loss (thus maximizing it). This reverse direction is combined with the direction suggested by the minimization of the label predictor's loss (as G_f and G_y are connected sequentially in a standard way). Overall, SGD training makes the features \mathbf{f} discriminative (good for predicting y), while trying to mix the the distributions $S(\mathbf{f})$ and $T(\mathbf{f})$ as much as possible. The resulting stochastic process can be seen as an example of *adversarial learning* and is reminiscent of adversarial generative networks [2].

Further outlook

Supervised deep learning methods are highly-successful across many applications. Yet training such models require lots of labeled data. Training them on surrogate data will therefore remain an important avenue for research. Unsupervised deep domain adaptation is becoming of particular interest for computer vision, since we almost always have some source of surrogate labeled data (the two most notable sources being Internet images and computer graphics).

The initial hope was that deep architectures will turn out to be invariant to domain shifts, yet this has not proven to be the case. On the one hand the networks show impressive ability to build invariance to some nuisance parameters towards higher level layers and thus mitigate the domain shift. On the other hand, the sheer number of parameters within modern deep architectures means that it is easier for deep models to overfit the peculiarities of a certain domain.

It is no wonder that several groups including ours started working in parallel on the unsupervised deep domain adaptation, i.e. training on labeled surrogate data and unlabeled target domain data (e.g. [1, 3, 4, 5]).

Overall, the goal seems to be to learn deep architectures where bottom layers are domain/modality specific with a gradually reducing specificity, middle layers are domain-invariant and task-unspecific, and then top layers are task specific (and class-specific). Parameters of the bottom layers of such networks can be either shared between domains or be different across domains.

References

- 1 Yaroslav Ganin and Victor Lempitsky, Unsupervised Domain Adaptation by Backpropagation, CoRR abs/1409.7495 (2014)
- 2 Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, Yoshua Bengio, Generative Adversarial Nets, NIPS 2014: 2672-2680
- 3 Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, Trevor Darrell: Deep Domain Confusion: Maximizing for Domain Invariance. CoRR abs/1412.3474 (2014)

- 4 Qiang Chen et al., Deep Domain Adaptation for Prediction of Fine-Grained Clothing Attributes. CVPR 2015
- 5 Mingsheng Long, Jianmin Wang: Learning Transferable Features with Deep Adaptation Networks. CoRR abs/1502.02791 (2015)

3.11 Feature Learning in a Probit Model with Correlated Noise

Stephan Mandt (Institute for Data Sciences and Engineering, Columbia University, US)

License  Creative Commons BY 3.0 Unported license
© Stephan Mandt

A large class of problems in statistical genetics amounts to finding a sparse linear effect in a binary classification setup, such as finding a small set of genes that most strongly predict a disease. Very often, these signals are spurious and obfuscated by confounders such as age, ethnicity or population structure. Beyond statistical genetics, sparse estimation is a general problem in binary classification, and has wide applications in science and technology, including, among many others, neuroscience, medicine, text classification, credit scoring, and computer malware detection. In all of these applications, confounding of the sparse signal can have dramatic consequences such as false medical diagnoses or violations of financial regulations. There is a need for statistical methods for feature selection that are robust to these confounding influences.

The model

In my talk I showed that by generalizing the probit model in a way that it captures correlated label noise is a way to eliminating confounders in the linear effect. Consider the following model:

$$Y_i = \text{sign}(X_i^\top w + \epsilon_i), \quad \epsilon = (\epsilon_1, \dots, \epsilon_n)^\top \sim \mathcal{N}(0, \Sigma).$$

This is just the probit regression model with the addition of a covariance matrix for the label noises. By making the simplifying assumptions that all observed labels are 1 (this can be achieved by a linear transformation on the noise covariance and data matrix), the central computational problem amounts to optimizing the following objective function:

$$\mathcal{L}(w) = -\log \int_{\mathbb{R}_+^n} \mathcal{N}(\epsilon; X^\top w, \Sigma) d^n \epsilon + \lambda_0 \|w\|_1.$$

Here, the ℓ_1 regularizer enforces sparsity in w , which is what we want in feature learning. In the uncorrelated case, the above integral decomposes into a sum of one-dimensional integrals that can be efficiently computed, but in the presence of correlations, the integral is intractable. In my talk, I derived an approximate inference algorithm for this task.

Why correlated label noises?

The correlated probit model delivers two alternative explanations of the observed labels Y_i : one in terms of a sparse linear effect (this is what we are interested in), and another explanation in terms of correlated label noise. The correlated label noise says, roughly speaking, that data points X_i that are similar, will also have similar labels Y_i . Similarity is

expressed in terms of a set of known kernels K_i (e.g., based on side information) that are the building blocks of the covariance matrix

$$\Sigma = \lambda_1 \mathbf{I} + \sum_{i=2}^m \lambda_i K_i.$$

The coefficients λ_i are determined by cross-validation. Now, by conditioning on the labels, the linear effect and the noise distribution will become correlated; in other words, thinking Bayesian, the correlated noise will *explain away* parts of the observed labels. Therefore the sparse linear effect will try to fit only those labels that are hard to fit with a correlated noise distribution, but better to fit with a sparse linear effect. Including a noise covariance matrix is therefore a possible way to include effects into our model that we do *not* want to have an effect on the sparse signal of interest.

Summary

Removing confounders in classification and regression task is an active and highly relevant field of research. A challenge is to make these more complex models computationally tractable. Variational methods offer a promising path.

References

- 1 Stephan Mandt, Florian Wenzel, Shinichi Nakajima, John P. Cunningham, Christoph Lipert, and Marius Kloft. Sparse Estimation in a Correlated Probit Model. arXiv preprint, arxiv:1507.04777.

3.12 A Resampling Method for Importance Weight Estimation

Shinichi Nakajima (TU Berlin, DE)

License © Creative Commons BY 3.0 Unported license
© Shinichi Nakajima

Joint work of Panknin, Danny; Braun, Mikio; Müller, Klaus-Robert;

Under the covariate shift setting, accurate estimation of importance weight is a key step, and several methods have been proposed for this purpose. We consider a new resampling method for density ratio estimation between two distributions, and introduce our plan to show its usefulness in theory and experiment.

3.13 Not IID Data in Advertising

Francesco Orabona (Yahoo! Labs – New York, US)

License © Creative Commons BY 3.0 Unported license
© Francesco Orabona

Main reference F. Orabona, “Simultaneous Model Selection and Optimization through Parameter-free Stochastic Learning,” in Proc. of the 2014 Annual Conf. on Neural Information Processing Systems (NIPS’14), pp. 1116–1124, 2014; pre-print available as arXiv:1406.3816v1 [cs.LG].

URL <http://papers.nips.cc/paper/5503-simultaneous-model-selection-and-optimization-through-parameter-free-stochastic-learning>

URL <http://arxiv.org/abs/1406.3816v1>

We present the problem of click prediction and show what is the most common solution employed in industry to not-IID training data. Latest achievements in automatic parameter tuning for stochastic gradient descent are also shown.

3.14 The Benefit of Multitask Representation Learning

Massimiliano Pontil (University College London, GB)

License  Creative Commons BY 3.0 Unported license
© Massimiliano Pontil

Main reference A. Maurer, M. Pontil, B. Romera-Paredes, “The Benefit of Multitask Representation Learning,” arXiv:1505.06279v1 [stat.ML], 2015.

URL <http://arxiv.org/abs/1505.06279v1>

We discuss a general method to learn data representations from multiple tasks. We provide a justification for this method in both settings of multitask learning and learning-to-learn. The method is illustrated in detail in the special case of linear feature learning. Conditions on the theoretical advantage offered by multitask representation learning over independent tasks learning are established. In particular, focusing on the important example of halfspace learning, we derive the regime in which multitask representation learning is beneficial over independent task learning, as a function of the sample size, the number of tasks and the intrinsic data dimensionality. Other potential applications of our results include multitask feature learning in reproducing kernel Hilbert spaces and multilayer, deep networks.

3.15 Adaptive Lifelong Learning for Visual Recognition and Data Analysis

Erik Rodner (Friedrich Schiller University Jena, DE)

License  Creative Commons BY 3.0 Unported license
© Erik Rodner

Current and previous work

Whereas my studies focused on transfer learning with Gaussian process models [8] and random decision forests [9], my current main research topic is lifelong learning and adaptive scientific data analysis. In particular, I have worked on aspects of adaptation [6, 10] (model sharing, learning from different but related datasets), active learning [3, 7] (selecting unlabeled examples which are likely beneficial when being labeled by an annotator), novelty detection [1] (determining whether an example belongs to an unknown category), and fine-grained recognition [11, 4, 2] (discriminating very similar categories). Learning with non-iid. data has been always part of my research on domain adaptation, where I search for handy solutions applicable to some of the large-scale learning problems we have in vision and scientific data analysis. One example is the MMDT (max-margin domain transforms) method presented in [6], which jointly learns classifier parameters as well as a linear transformation that maps labeled examples of one dataset to the feature space of another but related labeled dataset. The method itself is a straightforward extension of standard one-vs-all SVMs and can be used in large-scale scenarios [10].

Recently, people have boosted the performance on nearly all vision datasets and tasks by using a feature representation learned with high complexity models (*e.g.*, CNNs) on large-scale datasets, such as ImageNet. This strategy can be seen as non-iid. learning with two related but different datasets (ImageNet and another vision data set). In a recent publication, we brought this concept to an extreme by using pre-trained CNN models for object part discovery [11].

Adaptive lifelong learning

I am currently developing an approach which allows for adapting to new input data and especially new tasks (set of categories) in a semi-supervised learning setting. First of all, think about the scenario where we have ImageNet $\tilde{\mathcal{D}} = (\tilde{\mathbf{X}}, \tilde{\mathbf{y}})$ (labels $\tilde{\mathbf{y}}$, input examples $\tilde{\mathbf{X}}$), from which we can learn quite a lot of object categories, and an unlabeled set of images $\mathcal{D} = \mathbf{X}$ acquired in a new environment/domain (*e.g.*, video sequence of your office). The goal is now to learn an object classifier for the new domain by exploiting the fact that the input examples are related but different and the set of categories for the new domain might also contain new categories not part of ImageNet (have you ever searched for *toothpaste* in ImageNet?), *i.e.*, the label space changed.

In particular, $\tilde{\mathcal{D}}$ is sampled from $p(y, \tilde{\mathbf{x}}|\tilde{\mathbf{q}})$ and the unlabeled set \mathcal{D} is sampled from $p(\mathbf{x}|\mathbf{q})$, where $\tilde{\mathbf{q}}$ and \mathbf{q} are parameters of the distributions and are assumed to be sampled from a world model $p(\tilde{\mathbf{q}}|\mathbf{Q})$ and $p(\mathbf{q}|\mathbf{Q})$. The goal is to find a model for $p(y|\mathbf{x}, \mathbf{q})$ by using both datasets \mathcal{D} and $\tilde{\mathcal{D}}$ and carefully coupling of the distributions through the world model. In summary, this is a learning framework that allows for adaptation of the label and the input space jointly. Furthermore, it can be extended to learning over time by assuming continuously changing distributions parameterized by \mathbf{q}_t .

Further challenges

In general, I am also interested in studying the effects current fine-tuning strategies have for adaptation. In contrast to vision research a few years ago, people make indirectly use of domain adaptation principles when fine-tuning is performed on models initially learned on other datasets. How can we control the degree of adaptation performed? Are there any theoretical results that might help us to select the parameters that should be fine-tuned and the ones that should be fixed to their initial value?

Furthermore, adapting to the right output space, a user might need and expect, will be extremely important in future in my opinion, especially for scientific data analysis where the goal is not always defined in advance. In vision, object detection methods can now detect thousands of categories and without focusing and re-focusing on the subset and the granularity of semantic information the user needs, we are likely not be able to make use of the results at all.

References

- 1 Paul Bodesheim, Alexander Freytag, Erik Rodner, Michael Kemmler, and Joachim Denzler. Kernel null space methods for novelty detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3374–3381, 2013.
- 2 Alexander Freytag, Erik Rodner, Trevor Darrell, and Joachim Denzler. Exemplar-specific patch features for fine-grained recognition. In *German Conference on Pattern Recognition (GCPR)*, pages 144–156, 2014.
- 3 Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision (ECCV)*, volume 8692 of *Lecture Notes in Computer Science*, pages 562–577, 2014.
- 4 Christoph Göring, Erik Rodner, Alexander Freytag, and Joachim Denzler. Nonparametric part transfer for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2489–2496, 2014.
- 5 Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. LSDA: Large scale detection through adaptation. In *Neural Information Processing Systems (NIPS)*, 2014.

- 6 Judy Hoffman, Erik Rodner, Jeff Donahue, Brian Kulis, and Kate Saenko. Asymmetric and category invariant feature transformations for domain adaptation. *International Journal of Computer Vision (IJCV)*, 109(1-2):28–41, 2014.
- 7 Christoph Käding, Alexander Freytag, Erik Rodner, Paul Bodesheim, and Joachim Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- 8 Erik Rodner and Joachim Denzler. One-shot learning of object categories using dependent gaussian processes. In *Annual Symposium of the German Association for Pattern Recognition (DAGM)*, pages 232–241. Springer, 2010.
- 9 Erik Rodner and Joachim Denzler. Learning with few examples for binary and multi-class classification using regularization of randomized trees. *Pattern Recognition Letters*, 32(2):244–251, January 2011.
- 10 Erik Rodner, Judy Hoffman, Jeff Donahue, Trevor Darrell, and Kate Saenko. Transform-based domain adaptation for big data. In *NIPS Workshop on New Directions in Transfer and Multi-Task Learning*, 2013. abstract version of arXiv:1308.4200.
- 11 Marcel Simon, Erik Rodner, and Joachim Denzler. Part detector discovery in deep convolutional neural networks. In *Asian Conference on Computer Vision (ACCV)*, 2014.

3.16 Covariate Shift and Varying-Coefficient Models

Tobias Scheffer (University of Potsdam, DE)

License  Creative Commons BY 3.0 Unported license
© Tobias Scheffer

Joint work of Niels Landwehr, Matthias Bussas, Christoph Sawade, and Tobias Scheffer

The Past: Discriminative Learning of Importance Weights for Covariate Shift

Consider a data generation process in which there is a source variable $\sigma \in \{\text{train}, \text{test}\}$. Training instances are governed by $p(\mathbf{x}|\sigma = \text{train})$ whereas test instances are governed by a potentially different $p(\mathbf{x}|\sigma = \text{test})$. In either case, labels are created according to $p(y|\mathbf{x})$.

In order to minimize the regularized risk under the test distribution, one has to minimize

$$\sum_{i=1}^n \frac{p(\mathbf{x}|\sigma = \text{test})}{p(\mathbf{x}|\sigma = \text{train})} \ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \Omega(\mathbf{w}).$$

Estimating the training and test density functions [5] is unnecessarily difficult, because those are high-dimensional density functions and really only a scalar factor is needed for each instance. However, observe that, by simple arithmetics [1]:

$$\frac{p(\mathbf{x}|\sigma = \text{test})}{p(\mathbf{x}|\sigma = \text{train})} = \frac{p(\sigma = \text{train})}{p(\sigma = \text{test})} \left(\frac{1}{p(\sigma = \text{train}|\mathbf{x})} - 1 \right).$$

The density ratio can be written in terms of $p(\sigma = \text{train}|\mathbf{x})$ which can be estimated with a logistic regression model

$$p(\sigma = \text{train}|\mathbf{x}, \mathbf{v}) = \frac{1}{1 + \exp(\mathbf{v}^T \mathbf{x})}.$$

This model is trained using the training data as positive, and the test data as negative examples.

Over KLIEP [6], this method has the advantage that the optimization problems are more directly linked to minimizing the risk under the test distribution. Over kernel mean matching

[4] it has the advantage that the regularization parameter for model $f_{\mathbf{v}}$ can be tuned easily. Since it is trained on labeled data (with label σ), it can simply be tuned on held-out data.

The Future: Varying-Coefficient Models with Isotropic GP Priors

Consider problems with continuous task variables \mathbf{t} (e.g., time and space), regular attributes \mathbf{x} , and outputs y . Assume that $p_{\mathbf{t}}(y|\mathbf{x})$ changes smoothly in \mathbf{t} . For standard learning problems, parameters \mathbf{w} of a model $p(y|\mathbf{x}, \mathbf{w})$ are usually assumed to be governed by an isotropic Gaussian prior (hence ℓ_2 regularization of \mathbf{w}). Instead, let us assume that a function $\omega : \mathbf{t} \mapsto \mathbf{w}$ that generates task-specific parameters $\omega(\mathbf{t})$ of a model $p(y|\mathbf{x}, \omega(\mathbf{t}))$ is governed by an isotropic Gaussian Process prior.

The Gaussian Process couples $p(y|\mathbf{x}, \omega(\mathbf{t}))$ for different values of \mathbf{t} . A constant $\omega(\mathbf{t})$ corresponds to an *iid* model; generally, ω allows the model to change smoothly in \mathbf{t} .

“Theorem”. Let $\mathbf{X}, \mathbf{T}, \mathbf{y}$ be the training data and $\mathbf{x}^*, \mathbf{t}^*$ a test instance for which y^* has to be inferred. The predictive distribution $p(y^*|\mathbf{X}, \mathbf{y}, \mathbf{T}, \mathbf{x}^*, \mathbf{t}^*)$ of the above model is equal to the predictive distribution of a standard Gaussian process that uses concatenated attribute vectors (\mathbf{x}, \mathbf{t}) and product kernel $k((\mathbf{x}_i, \mathbf{t}_i), (\mathbf{x}_j, \mathbf{t}_j)) = k(\mathbf{x}_i, \mathbf{x}_j)k(\mathbf{t}_i, \mathbf{t}_j)$.

The theorem shows that Bayesian inference for varying-coefficient models can be done in $O(n^3 + dn)$ in the dual instead of in $O(n^3 d^3)$ [3] for n observations and d attributes. It also makes assumptions explicit that justify the use of products of task and instance kernels [2]. The model works great for geospatial problems such as predicting rents or real estate prices.

Acknowledgment. This is joint work with Niels Landwehr, Matthias Bussas, and Christoph Sawade.

References

- 1 S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *ICML*, 2007.
- 2 E. V. Bonilla, F. V. Agakov, and C. K. I. Williams. Kernel multi-task learning using task-specific features. In *AISTATS*, 2007.
- 3 A. Gelfand, H. Kim, C. Sirmans, and S. Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462), 2003.
- 4 J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Sample sel. bias by unlabeled data. In *NIPS*, 2007.
- 5 Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.
- 6 M. Sugiyama, T. Suzuki, S. Nakajima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Ann. of the Inst. of Stat. Math.*, 60(4), 2008.

3.17 Kernel Hypothesis Tests on Dependent Data

Dino Sejdinovic (University of Oxford, GB)

License © Creative Commons BY 3.0 Unported license
© Dino Sejdinovic

Joint work of Chwialkowski, Kacper; Sejdinovic, Dino; Gretton, Arthur

Main reference K.P. Chwialkowski, D. Sejdinovic, A. Gretton, “A wild bootstrap for degenerate kernel tests,” in Proc. of the 2014 Annual Conf. on Neural Information Processing Systems (NIPS’14), pp. 3608–3616, 2014.

URL <http://papers.nips.cc/paper/5452-a-wild-bootstrap-for-degenerate-kernel-tests>

Statistical tests based on embeddings of probability distributions into reproducing kernel Hilbert spaces have been applied in many contexts, including two sample testing [6], tests of independence [5, 1], tests of conditional independence [4, 10], and tests for higher order (Lancaster) interactions [8].

For these tests, consistency is guaranteed if and only if the observations are independent and identically distributed. Much real-world data fails to satisfy the i.i.d. assumption: audio signals, EEG recordings, text documents, financial time series, and samples obtained when running Markov Chain Monte Carlo (MCMC), all show significant temporal dependence patterns. The asymptotic behaviour of kernel test statistics becomes quite different when temporal dependencies exist – the difference in their asymptotic null distributions has important implications in practice: the permutation-based tests return an elevated number of false positives.

An alternative estimate of the null distribution for the problem of independence testing was proposed in [2] (where one signal is repeatedly *shifted* relative to the other). There is, however, no obvious way to generalise this approach to other testing contexts. For instance, we might have two time series, with the goal of comparing their marginal distribution. In [3], it was shown that an external randomization with *wild bootstrap* [7] may be applied to simulate from the null distribution for *all* kernel hypothesis tests for which V -statistics are employed, and not just for independence tests. This result has a potential to lead to a powerful set of model checking and MCMC diagnostic tools – where a nonparametric test can be constructed whether a Markov chain has reached its stationary distribution using Maximum Mean Discrepancy (MMD) [6] as a test statistic, similarly as in [9]. While a permutation-based test of whether the sampler has converged leads to too many rejections of the null hypothesis due to chain dependence (implying that one requires heavily thinned chains, which is wasteful of samples and computationally burdensome), the wild bootstrap approach can be applied directly on chains and is demonstrated to attain a desired number of false positives in [3].

Future Work

Consistency of the above procedures requires strong mixing conditions on the time series at hand. Moreover, the wild bootstrap procedure has a tuning parameter which requires some knowledge of the mixing properties in order to be properly calibrated. Finally, the interplay between the kernel choice and the test performance in the case of dependent data is not well understood. What are the inherent tradeoffs when trying to learn such tuning parameters on a held out portion of the data before performing a test? Moreover, many outstanding practical considerations arise in the application of tests to MCMC diagnostics. When to perform a test? Can tuning parameters be learned on the fly?

Acknowledgments. This is joint work with Kacper Chwialkowski and Arthur Gretton.

References

- 1 Besserve, M., Logothetis, N.K., and Schölkopf, B. Statistical analysis of coupled time series with kernel cross-spectral density operators. In *NIPS*. 2013.
- 2 Chwialkowski, K. and Gretton, A. A kernel independence test for random processes. In *ICML*, 2014.
- 3 Chwialkowski, K., Sejdinovic, D., and Gretton, A. A wild bootstrap for degenerate kernel tests. In *NIPS*, 2014.
- 4 Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *NIPS*, volume 20, pp. 489–496, 2007.
- 5 Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. A kernel statistical test of independence. In *NIPS*, volume 20, pp. 585–592, 2007.
- 6 Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., and Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.
- 7 Leucht, A. and Neumann, M.H. Dependent wild bootstrap for degenerate U- and V-statistics. *J. Multivar. Anal.*, 117:257–280, 2013.
- 8 Sejdinovic, D., Gretton, A., and Bergsma, W. A kernel test for three-variable interactions. In *NIPS*, pp. 1124–1132, 2013.
- 9 Sejdinovic, D., Strathmann, H., Lomeli Garcia, M., Andrieu, C., and Gretton, A. Kernel Adaptive Metropolis-Hastings. In *ICML*, 2014.
- 10 Zhang, K., Peters, J., Janzing, D., B., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. In *UAI*, 2011.

3.18 Zero-shot learning via synthesized classifiers

Fei Sha (University of Southern California – Los Angeles, US)

License © Creative Commons BY 3.0 Unported license
© Fei Sha

Joint work of Sha, Fei; Chao, Weilun; Changpinyo, Soravit; Gong, Boqing

Real-world objects have a long-tailed distribution, making it difficult to collect labeled images of rare objects for visual object recognition. One appealing way to address this problem is zero-shot learning. We propose a unified framework based on the key insight that the classifiers of semantically similar objects can be constructed from a set of *base* classifiers of “phantom” classes. In sharp contrast to previous work, the classifiers of both seen and unseen objects are synthesized from the base classifiers, enabling us to effectively learn the bases using the labeled data of the seen classes and then readily apply them to synthesizing the classifiers of unseen classes. We further consider a *generalized* zero-shot learning setting, in which the test phase is a multi-way classification problem over both seen and unseen classes. This generalized case reflects more closely how test data are distributed in real applications, leading to a more challenging task. We demonstrate superior performance of our approach over the state of the art for (generalized) zero-shot learning on two benchmark datasets.

I would like to acknowledge the beneficial discussions with Prof. Christoph Lampert (IST, Austria) at the Dagstuhl Seminar, in particular, pointers to his earlier work on generalized zero-shot learning.

3.19 A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems with an Application to Learning

Ingo Steinwart (Universität Stuttgart, DE)

License © Creative Commons BY 3.0 Unported license
© Ingo Steinwart

Joint work of Hang, Hanyuan; Steinwart, Ingo

Main reference H. Hang, I. Steinwart, “A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems with an Application to Learning,” arXiv:1501.03059v1 [math.PR], 2015.

URL <http://arxiv.org/abs/1501.03059v1>

We establish a Bernstein-type inequality for a class of stochastic processes that include the classical geometrically ϕ -mixing processes, Rio’s generalization of these processes, as well as many time-discrete dynamical systems. Modulo a logarithmic factor and some constants, our Bernstein-type inequality coincides with the classical Bernstein inequality for i.i.d. data. We further use this new Bernstein-type inequality to derive an oracle inequality for generic regularized empirical risk minimization algorithms and data generated by such processes. Applying this oracle inequality to support vector machines using the Gaussian kernels for both least squares and quantile regression, it turns out that the resulting learning rates match, up to some arbitrarily small extra term in the exponent, the optimal rates for i.i.d. processes.

3.20 Sampling without replacement: direct approach vs. reduction to i.i.d.

Ilya Tolstikhin (MPI for Intelligent Systems – Tübingen, DE)

License © Creative Commons BY 3.0 Unported license
© Ilya Tolstikhin

Joint work of Tolstikhin, Ilya; Blanchard, Gilles; Kloft, Marius

Main reference I. O. Tolstikhin, G. Blanchard, M. Kloft, “Localized complexities for transductive learning,” in Proc. of the 27th Conf. on Learning Theory (COLT’14), pp.857–884, JMLR.org, 2014.

URL <http://jmlr.org/proceedings/papers/v35/tolstikhin14.html>

We consider two closely related questions: (1) general properties of random variables sampled *without* replacement from arbitrary finite domains and (2) risk bounds in transductive learning, which is a particular setting of statistical learning theory introduced by V. Vapnik.

Formally, let $\mathcal{C} = \{c_1, \dots, c_N\}$ be some fixed finite *population*. Let Z_1, \dots, Z_n be sampled uniformly *without replacement* from \mathcal{C} for $n \leq N$. independent. which may be more useful depending on situations: n of them and then take the first subset. Random variables sampled without replacement naturally appear in many modern applications of statistics, probability, and machine learning. First example which comes to mind is cross-validation, where sample is randomly partitioned into training and validation subsets. Other examples include matrix completion problems, various iterative stochastic algorithms like stochastic gradient descent, low-rank matrix factorization problems, and many others.

Arguably, one of the most useful tools when it comes to analysis of stochastic procedures are *concentration inequalities*, which control a deviation of random variables from their expected values with high probability. Generally one would like to upper bound tail probabilities $\mathbb{P}\{\xi - E[\xi] > t\}$ or $\mathbb{P}\{E[\xi] - \xi > t\}$ for $t > 0$ and $\xi := f(X_1, \dots, X_n)$, where X_1, \dots, X_n are random variables taking values in domain X and $f: X^n \rightarrow \mathbb{R}$. The case when X_1, \dots, X_n are independent is very well studied and many useful results are available, including Hoeffding’s and Bernstein’s inequalities for sums of independent real-valued random variables and McDiarmid’s inequality for functions f with bounded differences. However,

when random variables are sampled without replacement $\xi := f(Z_1, \dots, Z_n)$ new techniques are needed.

First results in this direction were derived by Hoeffding, who showed that classic inequalities for sums mentioned above also hold for $\xi := \sum_{i=1}^n Z_i$. This result was based on the elegant *reduction* of the sampling without replacement scheme to the i.i.d. setting. Later results showed that a direct approach can be tighter than the reduction: using a *martingale technique* Serfling derived an improved version of Hoeffding's inequality for $\xi := \sum_{i=1}^n Z_i$, containing additional factor $\frac{N-n+1}{N}$ which decreases as $n \rightarrow N$. The same technique was later used to derive versions of Bernstein's and McDiarmid's inequalities for sampling without replacement, which improve upon the i.i.d. counterparts in the similar way.

In transductive learning, a learner observes n labeled training points together with u unlabeled test points with the final goal of giving correct answers for the test points. This process can be modeled using sampling without replacement described above, with fixed population of N input-output pairs $\mathcal{C} := \{(X_i, Y_i)\}_{i=1}^N$, random labeled training sample $S_n := \{Z_1, \dots, Z_n\}$, and unlabeled test sample X_u containing $u = N - n$ inputs of remaining elements $S_u := \mathcal{C} \setminus S_n$. Usually the learner fixes a class of predictors \mathcal{H} and a bounded loss function ℓ and seeks for an optimal predictor h_u^* minimizing an average test loss $\text{err}(h, S_u)$ over \mathcal{H} . However, labels of the test objects are unknown, and the learner resorts to \hat{h}_n which minimizes an empirical loss $\text{err}(h, S_n)$ over \mathcal{H} . The main question is: how large can be the *excess risk* $\text{err}(\hat{h}_n, S_u) - \text{err}(h_u^*, S_u)$? The excess risk can be upper bounded in a standard way by uniform deviations of risks computed on two disjoint finite samples $Q_n := \sup_{h \in \mathcal{H}} |\text{err}(h, S_u) - \text{err}(h, S_n)|$. Note that this construction naturally appears as a middle step in proofs of standard i.i.d. risk bounds as a result of symmetrization or the so-called *double-sample trick*. Since Q_n is a function of the random training set S_n , we can apply concentration inequalities for sampling without replacement in order to upper bound it using $E[Q_n]$. This can be done using a version of McDiarmid's inequality or more powerful versions of Talagrand's inequality for sampling without replacement, which were recently derived in [1].

It was also shown in [1] using Hoeffding's reduction trick that $E[Q_n]$ is upper bounded by $E[\tilde{Q}_n]$, where \tilde{Q}_n is a supremum of the standard i.i.d. empirical process. Using well-known *symmetrization inequalities* one can further upper bound $E[\tilde{Q}_n]$ (and thus $E[Q_n]$) with *Rademacher complexity* of the class \mathcal{H} . Together with concentration argument this shows that most of the i.i.d. risk bounds also hold in the transductive learning setting. However, we would like to argue that this reduction to i.i.d. setting can give suboptimal results compared to direct analysis of $E[Q_n]$ (in the same way as Hoeffding's reduction trick leads to suboptimal inequalities compared to the direct martingale technique).

We introduce a new complexity measure for transductive learning called *permutational Rademacher complexity* (PRC), which is similar to the standard Rademacher complexity. The only difference is that in PRC ± 1 signs are obtained using random permutation of a sequence containing equal number of “-1” and “+1”, while in the Rademacher complexity signs are sampled i.i.d. We provide the preliminary results on PRC, including a novel symmetrization inequality, which shows that $E[Q_n]$ is upper bounded by PRC.

References

- 1 Tolstikhin, I., Blanchard, G., Kloft, M.: Localized complexities for transductive learning. In: COLT 2014, pp. 857–884 (2014)

3.21 Active Learning for Domain Adaptation

Ruth Urner (*MPI for Intelligent Systems – Tübingen, DE*)

Joint work of Christopher Berlind and Ruth Urner

License  Creative Commons BY 3.0 Unported license

© Ruth Urner

Main reference C. Berlind, R. Urner, “Active Nearest Neighbors in Changing Environments,” in Proc. of the 32nd Int’l Conf. on Machine Learning (ICML’15), pp. 1870–1879, JMLR.org, 2015.

URL <http://jmlr.org/proceedings/papers/v37/berlind15.html>

While classic machine learning paradigms assume training and test data are generated from the same process, domain adaptation addresses the more realistic setting in which the learner has large quantities of labeled data from some *source* task but limited or no labeled data from the *target* task it is attempting to learn.

In the paper, we give the first formal analysis showing that using *active learning for domain adaptation* yields a way to address the challenges inherent in this scenario. As is common, we assume that the learner receives *labeled data* from the source task and *unlabeled data* from the target task. In our model, the learner can make a small number of queries for labels of target examples. Now the goal is to accurately learn a classifier for the target task while making as few label requests as possible.

We propose a simple nonparametric algorithm, ANDA, that combines an active nearest neighbor querying strategy with nearest neighbor prediction. ANDA receives a labeled sample from the source distribution and an unlabeled sample from the target task. It first actively selects a subset of the target data to be labeled based on the amount of source data among the k' nearest neighbors of each target example. Then it outputs a k -nearest neighbor classifier on the combined source and target labeled data.

We prove that ANDA enjoys strong performance guarantees on both the risk of the resulting classifier and the number of queries ANDA will make. Simply put, ANDA is guaranteed to make enough queries to be consistent but will not make unnecessary ones.

4 Working Groups, Presentations, and Panel Discussion

Working groups were an essential part of the seminar and have been integrated in the schedule in two versions: (1) discussion groups always directly after a presentation session and (2) working groups on Thursday during a longer time slot with topics voted for by the participants in a pseudo-random fashion. Especially the discussion groups directly after presentations led to interesting questions and comments by all participants. Although time was limited, results from the groups were summarized and supported a very interactive atmosphere of the seminar. The talks of the seminar had three different lengths: (1) longer keynotes for vision, algorithms, and computational biology for 25 minutes, (2) ongoing research talks for 12 minutes, and (3) quick presentations for just 3 minutes. This mix allowed a presentation for every participant and the quick presentations often led to interesting discussions in the evening.

The seminar ended with a panel discussion in the garden with Fei Sha, Shai-Ben David, and Oliver Stegle on the topic of open problems and upcoming research challenges in the area of non-i.i.d. learning. The topic quickly shifted towards recent advances in deep learning and how they are currently affecting the methodology used for non-i.i.d. learning. Especially for computational biology topics, the lack of large-scale training data was mentioned as the main obstacle for using these techniques. The panel ended with a summary of the seminar and a feedback to the organizers about its structure.

Participants

- Shai Ben-David
University of Waterloo, CA
- Gilles Blanchard
Universität Potsdam, DE
- Trevor Darrell
University of California –
Berkeley, US
- Joachim Denzler
Universität Jena, DE
- Philipp Drewe
Max-Delbrück-Centrum, DE
- Mario Fritz
MPI für Informatik –
Saarbrücken, DE
- Judy Hoffman
University of California –
Berkeley, US
- Josef Kittler
University of Surrey, GB
- Marius Kloft
HU Berlin, DE
- Brian Kulis
Ohio State University –
Columbus, US
- Christoph H. Lampert
IST Austria –
Klosterneuburg, AT
- Soeren Laue
Universität Jena, DE
- Alessandro Lazaric
INRIA – University of Lille 1, FR
- Victor Lempitsky
Skoltech – Skolkovo, RU
- Christoph Lippert
Los Angeles, US
- Stephan Mandt
Columbia University, US
- Shin Nakajima
TU Berlin, DE
- Francesco Orabona
Yahoo! Labs – New York, US
- Massimiliano Pontil
University College London, GB
- Gunnar Rätsch
Memorial Sloan-Kettering Cancer
Center – New York, US
- Erik Rodner
Universität Jena, DE
- Kate Saenko
University of Massachusetts –
Lowell, US
- Tobias Scheffer
Universität Potsdam, DE
- Dino Sejdinovic
University of Oxford, GB
- Fei Sha
University of Southern
California – Los Angeles, US
- Oliver Stegle
European Bioinformatics
Institute – Cambridge, GB
- Ingo Steinwart
Universität Stuttgart, DE
- Ilya Tolstikhin
MPI für Intelligente Systeme –
Tübingen, DE
- Ruth Urner
MPI für Intelligente Systeme –
Tübingen, DE
- Alexander Zimin
IST Austria –
Klosterneuburg, AT



Advanced Stencil-Code Engineering

Edited by

Christian Lengauer¹, Matthias Bolten², Robert D. Falgout³, and
Olaf Schenk⁴

- 1 Universität Passau, DE, christian.lengauer@uni-passau.de
- 2 Bergische Universität Wuppertal, DE, bolten@math.uni-wuppertal.de
- 3 Lawrence Livermore National Lab., US, falgout2@llnl.gov
- 4 University of Lugano, CH, olaf.schenk@usi.ch

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15161 “Advanced Stencil-Code Engineering”. The seminar was hosted by the DFG project with the same name (ExaStencils for short) in the DFG priority programme “Software for Exascale Computing” (SPPEXA). It brought together experts from mathematics, computer science and applications to explore the challenges of very high performance and massive parallelism in solving partial differential equations. Its aim was to lay the basis for a new interdisciplinary research community on high-performance stencil codes.

Seminar April 12–17, 2015 – <http://www.dagstuhl.de/15161>

1998 ACM Subject Classification C.4 Performance of Systems, D.1.3 Concurrent Programming, D.2.6 Programming Environments, D.3.3 Language Constructs and Features, G.1.8 Partial Differential Equations, G.4 Mathematical Software

Keywords and phrases Code generation, domain-specific languages, exascale computing, high-performance computing, massive parallelism, multigrid, partial differential equations, program optimization, program parallelization, stencil codes

Digital Object Identifier 10.4230/DagRep.5.4.56

1 Executive Summary

Christian Lengauer

Matthias Bolten

Robert D. Falgout

Olaf Schenk

License  Creative Commons BY 3.0 Unported license
© Christian Lengauer, Matthias Bolten, Robert D. Falgout, and Olaf Schenk

Stencil Codes

Stencil codes are compute-intensive algorithms, in which data points arranged in a large grid are being recomputed repeatedly from the values of data points in a predefined neighborhood. This fixed neighborhood pattern is called a stencil. Stencil codes see wide-spread use in computing the discrete solutions of partial differential equations and systems composed of such equations. Connected to the implementation of stencil codes is the use of efficient solver technology, i.e., iterative solvers that rely on the application of a stencil and that provide good convergence properties like multigrid methods. Major application areas are the natural sciences and engineering. Although, in many of these applications, unstructured



Except where otherwise noted, content of this report is licensed
under a Creative Commons BY 3.0 Unported license

Advanced Stencil-Code Engineering, *Dagstuhl Reports*, Vol. 5, Issue 4, pp. 56–75

Editors: Christian Lengauer, Matthias Bolten, Robert D. Falgout, and Olaf Schenk



DAGSTUHL
REPORTS

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

adaptive discretizations are employed for an efficient use of exascale supercomputers whose architectures possibly include accelerators or are of a heterogeneous nature, the use of structured discretizations and, thus, stencil codes has turned out to be helpful.

Stencil codes come in large varieties: there are many thousands! Deriving each of them individually, even if by code modification from one another, is not practical. The goal of the seminar is to raise the level of abstraction for application programmers significantly and to support this raise with an automated software technology that generates highly efficient massively parallel implementations which are tuned to the specific problem at hand and the execution platform used.

Research Challenges

Stencil codes are algorithms with a pleasantly high regularity: the data structures are higher-dimensional grids and the computations follow a static, locally contained dependence pattern and are typically arranged in nested loops with linearly affine bounds. This invites massive parallelism and raises the hope for easily achieved high performance. However, serious challenges remain:

- Because of the large numbers and varieties of stencil code implementations, deriving each of them individually, even if by code modification from one another, is not practical. Not even the use of program libraries is practical; instead, a domain-specific metaprogramming approach is needed.
- Reaching petascale to exascale execution speed is a challenge in the frequently used so-called multigrid algorithms, which work on a hierarchy of increasingly larger grids. The coarse grids in the upper part of the hierarchy are too small for massive parallelism.
- Efficiency, i.e., a high ratio of speedup to the degree of parallelism, is impaired by the low mathematical density, i.e., the low ratio of computation steps to data transfers of stencil codes.
- An inappropriate use of the execution platform may act as a performance brake.

Stencil-code engineering has received increased attention in the last few years, which is evidenced by the appearance of a number of stencil-code programming languages and frameworks. To reach the highest possible execution speed and to conserve hardware resources and energy, the stencil code must be tuned cleverly to the specific application problem at hand and the execution platform used. One approach that could be followed has been demonstrated by the previous U.S. project SPIRAL, whose target was the domain of linear transforms: domain-specific optimization at several levels of abstraction – from the mathematical equations over an abstract, domain-specific program and, in further steps, to the actual target code on the execution platform used. At each level, one makes aggressive use of knowledge of the problem and platform and employs up-to-date, automated software technology suitable for that level.

Questions and Issues Addressed

The charter of the seminar was to foster international cooperation in the development of a radically new, automatic, optimizing software technology for the effective and flexible exploitation of massively parallel architectures for dedicated, well delineated problem domains.

The central approaches in achieving this technology are:

- the aggressive use of domain knowledge for optimization at different levels of abstraction
- the exploitation of commonalities and variabilities in application codes via product-line technology and domain engineering

- the use of powerful models for program optimization, like the polyhedron model for loop parallelization and feature-orientation for software product lines

The application domain investigated in the seminar was stencil codes. It is envisaged that the approach can be ported to other well delineated domains – of course, with the substitution of suitable domain-specific content.

Among the issues discussed were:

- What are suitable abstraction, modularization, composition and generation mechanisms for stencil codes?
- What are the appropriate language features of a domain-specific language for stencil codes?
- What are the commonalities and variabilities of stencil codes?
- What are the computational performance barriers, especially, of multigrid methods using stencils and how can they be overcome?
- What are the performance barriers caused by data exchanges and how can they be overcome? How can communication be avoided in multilevel algorithms?
- What are the roles of nested loops and divide-and-conquer recursions in stencil codes?
- How can other solvers and preconditioners benefit from autotuned stencil codes?
- What role should techniques like autotuning and machine learning play in the optimization of stencil codes?
- What options of mapping stencil codes to a heterogeneous execution platform exist and how can an educated choice be made?
- Which techniques can be employed to make clever use of large-scale hybrid architectures, e.g., by the combination of multigrid with mathematical domain decomposition?

On the informatics side, one important role of the seminar was to inform the international stencils community about the techniques used in ExaStencils: software product lines, polyhedral loop optimization and architectural metaprogramming. Equally important was for ExaStencils members to learn about the experiences made with other techniques like divide-and-conquer, multicore optimization in parallel algorithms or autotuning. The application experts contributed to a realistic grounding of the research questions.

On the mathematics side, the seminar fostered the cooperation of experts in parallel solver technology with the groups from informatics to enable them to make use of the advanced techniques available. Further, different strategies for improving the scalability of iterative methods were discussed and the awareness of the opportunities and complexities of modern architectures in the numerical mathematics community was advanced.

2 Table of Contents

Executive Summary

Christian Lengauer, Matthias Bolten, Robert D. Falgout, and Olaf Schenk 56

Overview of Talks

From stencils to elliptic PDE solvers <i>Ulrich Rüde</i>	61
Maintaining performance in a general-purpose FEM code <i>Christian Engwer</i>	61
Optimization opportunities of stencils codes via analytic performance modeling <i>Georg Hager</i>	62
SPPEXA und ExaStencils <i>Christian Lengauer</i>	62
Local Fourier analysis for multigrid on semi-structured meshes <i>Carmen Rodrigo Cardiel</i>	62
Predicting the numerical performance of methods for evolutionary problems <i>Stephanie Friedhoff</i>	63
An extension of hypre's structured and semi-structured matrix classes <i>Ulrike Meyer-Yang</i>	64
The PyOP2 abstraction <i>Paul H. J. Kelly</i>	64
The Pochoir stencil compiler <i>Bradley Kuszmaul</i>	64
Formal synthesis of computational kernels <i>Franz Franchetti</i>	65
ExaSlang and the ExaStencils code generator <i>Christian Schmitt, Stefan Kronawitter, Sebastian Kuckuk</i>	65
Variability management in ExaStencils <i>Alexander Grebhahn</i>	66
From general-purpose to stencil DSL code <i>Armando Solar-Lezama</i>	66
STELLA: A domain-specific language for stencil computations <i>Carlos Osuna Escamilla</i>	67
Designing GridTools <i>Mauro Bianco</i>	67
Redesign of preconditioned Krylov methods around stencil compilers <i>Wim Vanroose</i>	68
PolyMage: High-performance compilation for heterogeneous stencils <i>Uday Bondhugula</i>	68
On the characterization of the data movement complexity of algorithms <i>P. Sadayappan</i>	69

The mathematics of ExaStencils <i>Hannah Rittich</i>	69
Stencils for hierarchical bases <i>Dirk Pflüger</i>	69
Mapping stencils to FPGAs and synthesizable accelerators <i>Louis-Noel Pouchet</i>	70
The stencil optimization framework MODESTO <i>Tobias Grosser</i>	70
Autotuning divide-and-conquer stencil computations <i>Ekanathan Palamadai Natarajan</i>	70
Things you can do with stencils <i>Gabriel Wittum</i>	71
Optimization of higher-order stencils <i>P. Sadayappan</i>	72
DSL for stencils in non-Newtonian fluids simulation? <i>Gundolf Haase</i>	72
Evening Discussion Sessions	72
Conclusions	74
Participants	75

3 Overview of Talks

40-min talk slots covered the programme until Thursday mid-afternoon. Many talks had multiple authors; in one, the presentation was shared by all authors. The latter part of the seminar was devoted to the planning of future collaborations. A list of talks follows in the order in which they were presented.

3.1 From stencils to elliptic PDE solvers

Ulrich Rüde (Friedrich-Alexander-Universität Erlangen-Nürnberg – Erlangen, DE)

License © Creative Commons BY 3.0 Unported license
© Ulrich Rüde

The talk addresses three aspects of stencil codes:

- What techniques can we use to speed up stencil codes? These are blocking and tiling techniques, but also memory layout transformations such as padding and multi-color-splits.
- What stencil codes should be considered? Here it is important to notice that the algorithms more often than not will need to accomplish a global exchange of data and that any attempt to avoid this will only result in inefficient algorithms. In other words: there is no way to avoid the complexities as they occur, e.g., in multigrid algorithms – the question is rather to find ways to implement such algorithmic structures as efficiently as possible.
- Where do we stand? Here the prototype HHG package shows that large systems with in excess of 10^{12} (a trillion) unknowns can be solved in a matter of minutes using highly optimized multigrid algorithms on parallel systems running up to a million parallel threads.

3.2 Maintaining performance in a general-purpose FEM code

Christian Engwer (Westfälische Wilhelms-Universität – Münster, DE)

License © Creative Commons BY 3.0 Unported license
© Christian Engwer
Joint work of P. Bastian, C. Engwer, J. Fahlke, S. Müthing

For solving partial differential equations the finite element method (FEM) is an attractive powerful tool. In many engineering applications, the FEM on unstructured meshes is used to account for the complicated geometry.

The text book stencil approaches gain high performance from their very simple predefined structure. FEM, on the other hand, gets the flexibility from its ability to deal with arbitrary unstructured meshes. To obtain the good performance of stencils and, at the same time, keep the flexibility of FEM, we adopt certain concepts from stencils. To achieve this, we introduce local structure – either by locally structured refinement or by higher-order methods.

A particular challenge arises from the fact that the PDE model is given in terms of user code and is executed in the inner most loop. We discuss how to restructure the interfaces to exploit the local structure and obtain a significant portion of peak performance while keeping the flexibility at the user level.

3.3 Optimization opportunities of stencils codes via analytic performance modeling

Georg Hager (Friedrich-Alexander-Universität Erlangen-Nürnberg – Erlangen, DE)

License  Creative Commons BY 3.0 Unported license
© Georg Hager

Much effort has been put into optimized implementations of stencil algorithms. Such activities are usually not guided by performance models that provide estimates of expected speedup. Understanding the performance properties and bottlenecks by performance modeling enables a clear view on promising optimization opportunities. We use the recently developed Execution-Cache-Memory (ECM) model to quantify the performance bottlenecks of stencil algorithms on a contemporary Intel processor. Single-core performance and scalability predictions for typical “corner-case” stencil loop kernels are given. Guided by the ECM model, we accurately quantify the significance of “layer conditions”, which are required to estimate the data traffic through the memory hierarchy, and study the impact of typical optimization approaches such as spatial blocking, strength reduction, and temporal blocking for their expected benefits. We also compare the ECM model to the widely known Roofline model and pinpoint the limitations of both. In an outlook, we demonstrate a simple tool that can automatically construct the Roofline and ECM models for streaming kernels (including stencils).

3.4 SPPEXA und ExaStencils

Christian Lengauer (Universität Passau – Passau, DE)

License  Creative Commons BY 3.0 Unported license
© Christian Lengauer

The DFG Priority Programme SPP 1648 “Software for Exascale Computing” (SPPEXA) is introduced briefly and one of its thirteen projects is sketched: “Advanced Stencil Code Engineering” (ExaStencils). The goals of the project are stated and justified, and the structure of the ExaStencils development framework is reviewed. Three further talks in the seminar provide details.

3.5 Local Fourier analysis for multigrid on semi-structured meshes

Carmen Rodrigo Cardiel (Universidad de Zaragoza – Zaragoza, ES)

License  Creative Commons BY 3.0 Unported license
© Carmen Rodrigo Cardiel

Joint work of A. Arrarás, F. J. Gaspar, B. Gmeiner, T. Gradl, F. J. Lisbona, L. Porter, C. Rodrigo Cardiel, U. Råde, P. Salinas

To approximate solutions of problems defined on complex domains, it is very common to apply a regular refinement to an unstructured input grid which fits the geometry of the domain. In this way, a hierarchy of globally unstructured grids with regular structured patches is generated. This kind of mesh is suitable for the use with geometric multigrid methods and allows us to use stencil-based data structures which reduce the memory requirements drastically.

In this setting, we are interested in the design of efficient geometric multigrid methods on such semi-structured triangular grids. To design these algorithms, a local Fourier analysis for non-orthogonal grids is used. This tool is based on the Fourier transform and provides very accurate predictions of the asymptotic convergence of geometric multigrid methods. It is a useful technique for the choice of the suitable components of your algorithm. This analysis is applied on each structured patch of the grid in order to choose the most efficient components on each block of the semi-structured grid.

This strategy was initially applied to linear finite-element discretizations of scalar partial differential equations in two dimensions, and was later extended and generalized to such discretizations for systems of PDEs, three-dimensional tetrahedral grids, high-order finite element discretizations, finite-volume cell-centered schemes, and even to time-dependent non-linear problems by combining the approach with splitting schemes in time.

Note that each of these extensions requires the design of specific smoothers appropriate for the problems encountered with the different discretizations and, most of the time, it is also a special approach of the local Fourier analysis is necessary.

3.6 Predicting the numerical performance of methods for evolutionary problems

Stephanie Friedhoff (University of Leuven – Leuven, BE)

License © Creative Commons BY 3.0 Unported license
© Stephanie Friedhoff

Joint work of S. Friedhoff, S. MacLachlan

With current trends in computer architectures leading towards systems with more, but not faster, processors, faster time to solution must come from greater parallelism. These trends particularly impact the numerical solution of the linear systems arising from the discretization of partial differential equations (PDEs) with evolutionary behavior, such as parabolic problems. The multigrid-reduction-in-time (MGRIT) algorithm is a truly multi-level approach to parallel-in-time integration, which directly uses an existing time propagator and, thus, can easily exploit substantially more computational resources than standard sequential time stepping. Multigrid waveform relaxation is another effective multigrid method on space-time grids for parabolic problems. However, a large gap exists between the theoretical analysis of these algorithms and their actual performance.

We present a generalization of the well-known local-mode (often local Fourier) analysis (LFA) approach. The proposed semi-algebraic mode analysis (SAMA) approach couples standard LFA with tractable numerical computation that accounts for the non-local character of operators in the class of evolutionary problems. We demonstrate that SAMA provides an advantage for parabolic problems, obtaining robust predictivity of performance independent of the length of the time domain – in sharp contrast to LFA, which only becomes predictive for extremely long time integration.

3.7 An extension of hypre’s structured and semi-structured matrix classes

Ulrike Meier Yang (LLNL – Livermore, US)

License © Creative Commons BY 3.0 Unported license

© Ulrike Meyer-Yang

Joint work of R. Falgout, U. Meyer-Yang

The hypre software library provides high-performance preconditioners and solvers for the solution of large sparse linear systems on massively parallel computers. One of its attractive features is the provision of conceptual interfaces, which include a structured interface, a semi-structured interface, and a traditional linear-algebra-based interface. The (semi-)structured interfaces are an alternative to the standard matrix-based interface that describes rows, columns, and coefficients of a matrix. Here, instead, matrices are described primarily in terms of stencils and logically structured grids. These interfaces give application users a more natural means for describing their linear systems, and provide access to methods such as structured multigrid solvers, which can take advantage of the additional information beyond just the matrix. Since current architecture trends are favoring regular compute patterns to achieve high performance, the ability to express structure has become much more important.

We describe a new structured-grid matrix class that supports rectangular matrices and constant coefficients and a semi-structured-grid matrix class that builds on the new structured-grid matrix. We anticipate that an efficient implementation of these new classes will lead to better performance of matrix kernels and algorithms on current and future architectures than hypre’s current matrix classes.

3.8 The PyOP2 abstraction

Paul H. J. Kelly (Imperial College – London, UK)

License © Creative Commons BY 3.0 Unported license

© Paul H. J. Kelly

Joint work of G. Bercea, C. Bertolli, C. Cantwell, M. Giles, G. Gorman, D. A. Ham, P. H. J. Kelly, C. Krieger, F. Luporini, G. R. Markall, M. Mills Strout, G. Mudalige, C. Olschanowsky, R. Ramanujam, F. Rathgeber, I. Reguly, G. Rokos, S. Sherwin

PyOP2 is a stencil-like abstraction for parallel loops over unstructured meshes. It is used as an intermediate representation in Firedrake, an automated system for the portable solution of partial differential equations using the finite element method. This talk explores some of the opportunities exposed by PyOP2, for unstructured and extruded meshes and for locality, parallelisation and vectorisation.

3.9 The Pochoir stencil compiler

Bradley Kuszmaul (Massachusetts Institute of Technology – Boston, MA, US)

License © Creative Commons BY 3.0 Unported license

© Bradley Kuszmaul

Joint work of R. A. Chowdhury, B. Kuszmaul, C. E. Leiserson, C.-K. Luk, Y. Tang

A stencil computation repeatedly updates each point of a d -dimensional grid as a function of itself and its near neighbors. Parallel cache-efficient stencil algorithms based on “trapezoidal

decompositions” are known, but most programmers find them difficult to write. The Pochoir stencil compiler allows a programmer to write a simple specification of a stencil in a domain-specific stencil language embedded in C++ which the Pochoir compiler then translates into high-performing Cilk code that employs an efficient parallel cache-oblivious algorithm. Pochoir supports general d -dimensional stencils and handles both periodic and aperiodic boundary conditions in one unified algorithm. The Pochoir system provides a C++ template library that allows the user’s stencil specification to be executed directly in C++ without the Pochoir compiler (albeit more slowly), which simplifies user debugging and greatly simplified the implementation of the Pochoir compiler itself. A host of stencil benchmarks run on a modern multicore machine demonstrates that Pochoir outperforms standard parallel-loop implementations, typically running 2–10 times faster. The algorithm behind Pochoir improves on prior cache-efficient algorithms on multidimensional grids by making “hyperspace” cuts, which yield asymptotically more parallelism for the same cache efficiency.

3.10 Formal synthesis of computational kernels

Franz Franchetti (Carnegie Mellon University – Pittsburgh, PA, US)

License © Creative Commons BY 3.0 Unported license
© Franz Franchetti

Joint work of research groups SPIRAL and HACMS

We address the question of how to map computational kernels automatically across a wide range of computing platforms to highly efficient code, and prove the correctness of the synthesized code. This addresses two fundamental problems that software developers are faced with: performance portability across the ever-changing landscape of parallel platforms, and verifiable correctness of sophisticated floating-point code. We have implemented this approach as part of the SPIRAL system where we have formalized a selection of computational kernels from the signal and image processing domain, software-defined radio, and robotic vehicle control.

3.11 ExaSlang and the ExaStencils code generator

Christian Schmitt (Friedrich-Alexander-Universität Erlangen-Nürnberg – Erlangen, DE)

Stefan Kronawitter (Universität Passau – Passau, DE)

Sebastian Kuckuk (Friedrich-Alexander-Universität Erlangen-Nürnberg – Erlangen, DE)

License © Creative Commons BY 3.0 Unported license
© Christian Schmitt, Stefan Kronawitter, Sebastian Kuckuk

Joint work of F. Hannig, H. Köstler, S. Kronawitter, S. Kuckuk, C. Lengauer, U. Rüde, C. Schmitt, J. Teich

Many problems in computational science and engineering involve elliptic partial differential equations and require the numerical solution of large, sparse (non-)linear systems of equations. Multigrid is known to be one of the most efficient methods for this purpose. However, the concrete multigrid algorithm and its implementation depend highly on the underlying problem and hardware.

Project ExaStencils aims at a compiler and underlying code generation framework capable of generating automatically highly parallel and highly efficient geometric multigrid solvers from a very abstract description, while selecting the most performant program composition.

In our presentation, we focus on the aspects of code generation, node-level performance optimization and parallelization. More specifically, we provide an insight into the different components of our compiler framework and introduce some of the polyhedral as well as the low-level optimizations which are applied automatically. Furthermore, we introduce our approach to domain partitioning as well as details on employed communication strategies. Finally, we present several experimental results, ranging from node-level performance improvements over a case study of adding generator support for FPGAs to weak-scaling results on JUQUEEN.

3.12 Variability management in ExaStencils

Alexander Grebhahn (Universität Passau – Passau, DE)

License  Creative Commons BY 3.0 Unported license
© Alexander Grebhahn

Joint work of S. Apel, A. Grebhahn, N. Siegmund

The automatic generation of multigrid solvers leads to the possibility of creating a high number of different variants that are optimal for a wide range of different hardware. However, identifying the optimal variant for a specific hardware is a challenging task due to the inherent variability of the solvers. To master this challenge, we created a machine-learning approach for the derivation of a performance-influence model. Such a model describes all relevant influences of configuration options and their interactions on the performance of all possible variants. To identify a performance-influence model, we use an iterative approach that relies on a number of measurements gathered, using several structured sampling heuristics. In a series of experiments, we demonstrated the feasibility of our approach in terms of accuracy of the derived models.

3.13 From general-purpose to stencil DSL code

Armando Solar-Lezama (Massachusetts Institute of Technology – Boston, MA, US)

License  Creative Commons BY 3.0 Unported license
© Armando Solar-Lezama

The talk describes a new technique to allow high-performance stencil DSLs to be used to optimize existing legacy code. The key idea is to use synthesis technology to derive a high-level specification from a block of low-level code implementing a stencil. This high-level specification can then be mechanically translated into a target DSL.

In addition to deriving the code, the technique also derives the invariants necessary to prove the equivalence between the code and the extracted specification. The talk presented some initial results that suggest that the technique can extract a specification from some non-trivial stencils, including some that have undergone some hand optimization.

3.14 STELLA: A domain-specific language for stencil computations

Carlos Osuna Escamilla (Eidgenössische Technische Hochschule – Zürich, CH)

License © Creative Commons BY 3.0 Unported license
© Carlos Osuna Escamilla

Joint work of M. Bianco, O. Fuhrer, T. Gysi, C. Osuna Escamilla, T. C. Schulthess

The dynamical core of many weather and climate simulation models are implemented as stencil methods solving partial differential equations (PDEs) on structured grids. STELLA has been developed as a domain-specific language, based on C++ template metaprogramming, for stencil codes on structured grids. The library abstracts the loop logic and parallelization of the stencils as well as other hardware-dependent optimizations like memory layout of data fields, loop and kernel fusion or software manages cache techniques. We show the use and performance of different parallelization algorithms within STELLA, like a parallel tridiagonal solver. These new parallelization modes increase the level of parallelism in GPUs for the algorithmic motifs used in COSMO, which improves the strong scaling behaviour and therefore time to solution on real use cases. A full rewrite of the COSMO dynamical core shows the usefulness of STELLA for production codes, when stencil computations often have to interoperate with other parts of the model using different programming models and even programming languages. Using STELLA, we achieved a speedup factor of 1.8x for CPUs and 5.8x for NVIDIA GPUs for the dynamical core of COSMO.

3.15 Designing GridTools

Mauro Bianco (Eidgenössische Technische Hochschule – Zürich, CH)

License © Creative Commons BY 3.0 Unported license
© Mauro Bianco

The complexity and diversity of contemporary computer architectures make the effort of developing portable and efficient monolithic scientific applications a challenging endeavor. An application may use different algorithmic motifs, which have different requirements and optimal solution strategies.

Previous experience, such as STELLA, showed that investing in generalizing high-level application-specific libraries for portions of an application, provides performance portability. Thanks to this, the application can exploit new energy-efficient architectures, thus enabling innovative solutions of scientific problems, like increasing the resolution of numerical simulations for weather forecast.

GridTools is a set of C++ generic APIs to encapsulate the main algorithmic motifs in grid applications, such as weather and climate simulations. Specifically, GridTools provides a DSL for stencils computations, plus other facilities for halo-exchange communications, boundary conditions treatment, etc. This allows an application to be specified at high level and, at the same time, take advantage of the diverse architectural features, such as, multiple address spaces in modern computing nodes.

By means of its clean design and concepts, and being written in a very well established programming language, the GridTools API set is engineered to be expanded and improved over time, providing production a quality implementation of its components.

3.16 Redesign of preconditioned Krylov methods around stencil compilers

Wim Vanroose (University of Antwerp – Antwerp, BE)

License  Creative Commons BY 3.0 Unported license
© Wim Vanroose

Joint work of S. Donfack, P. Ghysels, B. Reys, O. Schenk, W. Vanroose

The performance of preconditioned Krylov solvers is severely hampered by the limited memory bandwidth. Each of the building blocks of a multigrid preconditioned Krylov solver is an operation of low arithmetic intensity. We discuss how the algorithm can be organized using stencil compilers such the arithmetic intensity is raised, so we can benefit from SIMD operations.

3.17 PolyMage: High-performance compilation for heterogeneous stencils

Uday Bondhugula (Indian Institute of Science – Bangalore, IN)

License  Creative Commons BY 3.0 Unported license
© Uday Bondhugula

Joint work of U. Bondhugula, R. T. Mullapudi, V. Vasista

This talk presents the design and implementation of PolyMage, a domain-specific language and compiler for image processing pipelines. An image processing pipeline can be viewed as a graph of interconnected stages which process images successively. Each stage typically performs one of point-wise, stencil, reduction or data-dependent operations on image pixels. Individual stages in a pipeline typically exhibit abundant data parallelism that can be exploited with relative ease. However, the stages also require high memory bandwidth preventing effective utilization of parallelism available on modern architectures. For applications that demand high performance, the traditional options are to use optimized libraries like OpenCV or to optimize manually. While using libraries precludes optimization across library routines, manual optimization accounting for both parallelism and locality is very tedious.

The focus of our system, PolyMage, is on automatically generating high-performance implementations of image processing pipelines expressed in a high-level declarative language. Our optimization approach primarily relies on the transformation and code generation capabilities of the polyhedral compiler framework. To the best of our knowledge, this is the first model-driven compiler for image processing pipelines that performs complex fusion, tiling, and storage optimization automatically. Experimental results on a modern multicore system show that the performance achieved by our automatic approach is up to 1.81x better than that achieved through manual tuning in Halide, a state-of-the-art language and compiler for image processing pipelines. For a camera raw image processing pipeline, our performance is comparable to that of a hand-tuned implementation.

3.18 On the characterization of the data movement complexity of algorithms

P. Sadayappan (Ohio State University – Columbus, OH, US)

License © Creative Commons BY 3.0 Unported license

© P. Sadayappan

Joint work of V. Elango, L.-N. Pouchet, J. Ramanujam, F. Rastello, P. Sadayappan

Data movement costs represent a significant factor in determining the execution time and energy expenditure of algorithms on current/emerging computers. Hence, characterizing the data movement cost is becoming increasingly important. This talk introduces the problem of finding lower bounds on the data movement complexity of algorithms and summarizes recent progress along multiple directions: a new approach to lower bounds using graph min-cut, automated analysis of affine codes for parametric asymptotic characterization of data movement lower bounds, and algorithm-architecture co-design using lower bounds on data movement.

3.19 The mathematics of ExaStencils

Hannah Rittich (Bergische Universität Wuppertal – Wuppertal, DE)

License © Creative Commons BY 3.0 Unported license

© Hannah Rittich

Joint work of M. Bolten, K. Kahl, H. Rittich

Local Fourier analysis (LFA) is a well known tool for analyzing and predicting the convergence behavior of multigrid methods. Originally, LFA has been introduced to analyze operators with constant coefficients. However, for some problems this assumption is too restrictive. We present a generalization of LFA that allows to analyze more general operators. The coefficients of these operators may vary in a block structured way. This enables us to analyze complex operators like operators with jumping coefficients and block smoothers.

3.20 Stencils for hierarchical bases

Dirk Pflüger (Universität Stuttgart – Stuttgart, DE)

License © Creative Commons BY 3.0 Unported license

© Dirk Pflüger

The discretization and solution of higher-dimensional PDEs suffers the curse of dimensionality. In these settings, hierarchical approaches such as sparse grids can help to push the limit of the dimensionality that can be dealt with. Sparse grids are based on hierarchical basis functions with local support and a truncation of the resulting expansion into higher-dimensional increment spaces. However, a FEM approach leads to matrices that are no longer sparse and to bilinear forms that result in “hierarchical stencils”. Thus, solution techniques and data structures that are optimized for sparse linear systems cannot be applied any more. But algorithms for the matrix-vector multiplication in optimal (linear) complexity do exist, even though an explicit assembly of the corresponding matrix would lead to more non-zero entries. In this talk, we present the challenges that arise when employing hierarchical bases and discuss the current state of the art with respect to the solution of PDEs. Novel ideas from other fields are more than welcome.

3.21 Mapping stencils to FPGAs and synthesizable accelerators

Louis-Noel Pouchet (Ohio State University – Columbus, OH, US)

License  Creative Commons BY 3.0 Unported license
© Louis-Noel Pouchet

Stencils are recurring patterns in numerous application domains, ranging from image processing to PDE solving. The medical imaging domain leverages numerous algorithms using stencils, and achieving portable high performance for these codes requires an integrated approach, from application modeling to low-level compiler optimization and hardware design.

In this talk, I present an overview of the research at the Center for Domain-Specific Computing, which aims at accelerating medical imaging applications by combining domain-specific modeling of applications, domain- and pattern-specific optimizing compilers, and a customizable heterogeneous computing platform. Focus will be made on the high-level modeling of the application using macro-dataflow concepts, and domain-specific optimization for stencils on CPUs and FPGAs.

3.22 The stencil optimization framework MODESTO

Tobias Grosser (Eidgenössische Technische Hochschule – Zürich, CH)

License  Creative Commons BY 3.0 Unported license
© Tobias Grosser

Joint work of Tobias Grosser, Tobias Gysi, Torsten Hoefler

Code transformations, such as loop tiling and loop fusion, are of key importance for the efficient implementation of stencil computations. However, their direct application to a large code base is costly and severely impacts program maintainability. While recently introduced domain-specific languages facilitate the application of such transformations, they typically still require manual tuning or auto-tuning techniques to select the transformations that yield optimal performance. We introduce MODESTO, a model-driven stencil optimization framework, that for a stencil program suggests program transformations optimized for a given target architecture. Initially, we review and categorize data locality transformations for stencil programs and introduce a stencil algebra that allows the expression and enumeration of different stencil program implementation variants. Combining this algebra with a compile-time performance model, we show how to automatically tune stencil programs. We use our framework to model the STELLA library and optimize kernels used by the COSMO atmospheric model on multi-core and hybrid CPU-GPU architectures. Compared to naive and expert-tuned variants, the automatically tuned kernels attain a 2.0–3.1x and a 1.0–1.8x speedup, respectively.

3.23 Autotuning divide-and-conquer stencil computations

Ekanathan Palamadai Natarajan (Massachusetts Institute of Technology – Boston, MA, US)

License  Creative Commons BY 3.0 Unported license
© Ekanathan Palamadai Natarajan

Joint work of C. E. Leiserson, E. Palamadai Natarajan

Ztune is an application-specific autotuner for optimizing serial divide-and-conquer stencil computations modeled on the trapezoidal-decomposition algorithm due to Frigo and Strumpfen. Each recursive step of the algorithm divides a space-time hypertrapezoidal region, called a

“zoid”, into subzoids, or if the size of the zoid falls beneath a given threshold, executes a base case that loops across the space-time dimensions to compute the stencil at each point in the zoid. Ztune defines a search domain that generalizes this algorithm, where at each zoid in the recursion tree, a parameter is created that chooses whether to divide or perform the base case and, if divide, chooses the space-time dimension to be divided. Ztune searches this domain of possible choices to find the fastest plan for executing the stencil computation. Although Ztune, in principle, performs an exhaustive search of the search domain, it uses three properties – space-time equivalence, divide subsumption, and favored dimension – to prune the search domain. These three properties reduce the autotuning time by orders of magnitude without significantly sacrificing runtime.

We implemented Ztune to autotune the Trap algorithm used in the open-source Pochoir stencil compiler (disabling parallelism). We then compared the performance of the Ztuned code with that of Pochoir’s default code on nine application benchmarks across two machines with different hardware configurations. On these benchmarks, the Ztuned code ran 5%–8% faster (geometric mean) than Pochoir’s hand-optimized code.

We also compared Ztune with state-of-the-art heuristic autotuning. The sheer number of choice parameters in Ztune’s search domain, however, renders naive heuristic autotuning infeasible. Consequently, we used the open-source OpenTuner framework to implement a heuristic autotuner called Otune that optimizes only the size of the base case along each dimension. Whereas the time for Ztune to autotune each of the benchmarks could be measured in minutes, to achieve comparable results, the autotuning time for Otune was typically measured in hours or days. Surprisingly, for some benchmarks, Ztune actually autotuned faster than the time it takes to perform the stencil computation once.

3.24 Things you can do with stencils

Gabriel Wittum (Goethe-Universität – Frankfurt, DE)

License © Creative Commons BY 3.0 Unported license
© Gabriel Wittum

Numerical simulation with supercomputers has become one of the major topics in computational science. To promote modelling and simulation of complex problems, new strategies are needed allowing for the solution of large, complex model systems. Crucial issues for such strategies are reliability, efficiency, robustness, scalability, usability, and versatility.

After introducing some basic notions of stencil notation and calculus, we discuss what is necessary for computing with a fixed or almost fixed stencil. To demonstrate this, we discuss the computation of seiches of Lake Constance using a domain adapted but structured grid, which still yields an operator with almost fixed pattern.

We discuss advantages and disadvantages and show that a numerical approach combining adaptivity, parallelism and multigrid methods allows for extreme parallel scalability while still maintaining flexibility to adapt numerical methods to the problem is more general and allows to gain acceleration by factors of up to 10^6 using sensible adaptive numerics. These strategies are combined in the novel simulation system UG 4 (“Unstructured Grids”).

3.25 Optimization of higher-order stencils

P. Sadayappan (Ohio State University – Columbus, OH, US)

License  Creative Commons BY 3.0 Unported license
© P. Sadayappan

Joint work of M. Kong, L.-N. Pouchet, J. Ramanujam, F. Rastello, P. Sadayappan, K. Stock

It is well known that the associativity of operations like addition and multiplication offer opportunities for reordering of operations to enable better parallelization. We discuss a complementary use of associativity of operations: to improve data locality. We consider the optimization of high-order (or long-range) stencil computations. High-order stencil computations exhibit a much higher operational intensity (ratio of arithmetic operations to data moved from/to main-memory) than simple nearest-neighbor stencil operations. Although high-order stencil computations are not memory-bandwidth-bound due to high operational intensity, they nevertheless do not achieve much higher performance than low-order stencils. This is because of severe register pressure. We discuss an associative reordering strategy that interleaves computations targeting multiple neighboring grid sites and thereby significantly reduces register pressure. Experimental results demonstrate significant performance improvement through use of the associative reordering.

3.26 DSL for stencils in non-Newtonian fluids simulation?

Gundolf Haase (Karl-Franzens-Universität – Graz, AT)

License  Creative Commons BY 3.0 Unported license
© Gundolf Haase

Joint work of D. Vasco

The talk introduces the formulation of non-Newtonian fluids as a coupled system of non-linear PDEs. The included Navier-Stokes equations contain an additional non-Newtonian term originating from the non-linear shear-stress tensor. The non-linear system of equations in each time step is solved by the SIMPLE algorithm and the discretization is the 7-point stencil taking into account the staggered grid for the pressure and the temperature.

Several issues for DSLs in this context are discussed including the automatic generation of a geometric multigrid for the coupled equations of the linearized system of equations. The final goal would be a full approximation scheme (FAS) automatically generated for multicore CPUs/GPUs/Xeon Phi and MPI.

4 Evening Discussion Sessions

Stencil Codes: Walls, Challenges, Goals

The initial goal of the discussion was to pinpoint the problems which stand in the way of the scalability of stencil codes and the productivity of their engineering. Ulrich Rüde pointed out that when solving linear systems where the system matrix is described by a stencil using direct methods is not feasible as the inverse of such a matrix is usually almost dense. Using separators does not solve this problem either: because the Schur complement is usually dense, an iterative solution results in too many iterations, thus multigrid approaches are needed. While asymptotically multigrid is the fastest solver in many cases, in practice they

are sometimes not the best choice as the constants involved depend on the problem at hand and not only the algorithm determines complexity.

In many cases Krylov subspace methods are used, but Saday Sadayappan mentioned that Krylov subspace methods are inefficient with respect to data movement. Another issue is the missing algorithmic scalability, i.e., the dependence of the number of iterations on the problem size. This could be overcome by multigrid preconditioning but essentially this puts the hierarchy necessary to tackle the problem in the preconditioner.

Paul Kelly and Gabriel Wittum further explained that the performance of the method heavily depends on the problem, e.g., in compressible or incompressible flows or when dealing with parabolic or hyperbolic PDEs. Christian Engwer noted that the building blocks are the same, although as countered by Saday Sadayappan the applications are different.

Harald Köstler added that people from both the compiler community and applied mathematics are needed to solve a problem. Further, at the interface of both worlds, experts are needed. Saday Sadayappan replied that the computer science people only need to know about the loops. It was argued that the algorithmic choice heavily influences the time to solution but this could be handled by optimizing all different algorithmic options.

According to Franz Franchetti one should start with a short compact way to describe the algorithm and optimize from there. Harald Köstler added that actually a top-down approach starting from the model is the right approach. According to Christian Engwer the optimal way would be if mathematicians provide a bilinear form plus information about the mesh and that computer scientists optimize starting from this. Therefore a rich language is needed that stated by Armando Solar-Lezama should be high-level enough. Sven Apel added that a domain scoping is necessary and that abstraction hides complexity. In Franz Franchetti's opinion several layers of DSLs are needed. In the following discussion it was noted that designing a DSL targeting HPC is difficult as some concepts are not scalable or optimizable. Further, it was added that currently hand-written and generated codes are compared. Understanding the domain makes creating optimized implementations easier, resulting in a huge increase of productivity. The necessary domain-specific optimization needs semantic knowledge or should at least benefit from it stated Paul Kelly. Franz Franchetti closed the discussion by pointing out that one has to be able to express what is known.

Suitable Representations of Stencil Codes

One issue at the seminar was multi-layered domain-specific optimization. Three approaches were presented: FireDrake, SPIRAL and ExaStencils. Each one offers at the most abstract, a language of mathematical expressions in the considered domain and at the more concrete levels executable representations that include successively aspects of the computation and architecture.

Starting point of the discussion was a slide (slide 20) of Franz Franchetti's talk showing the SPIRAL languages at different levels of abstraction for the three domains that SPIRAL has handled in the past (linear transforms, linear algebra, ...). In his talk, Franchetti had left open whether such representations could be developed for multigrid solving.

The second-most abstract and first executable level in SPIRAL has a representation in point-free combinator style. This means that the equations are in the space of the functions themselves and not in the range of the functions. Syntactically it is distinguished by the absence of function arguments in the expressions. It was explored whether and how this could be achieved for multigrid. The issue of matrix-freedom was also discussed. Matrix-freedom

gets rid of the need to store intermediate matrices in a computation. The conclusion was that multigrid solving seems to lend itself well to the SPIRAL style of software engineering. This will be explored in further collaborations.

A discussion as to the sensibility of restricted domains ensued. The criticism voiced was that a domain-specific approach has little relevance because much – maybe, most – application software will not be covered by it. It was countered by two arguments: (1) it is unlikely that a general-purpose software technology can be as powerful as a domain-specific one and (2) if the domain has a wide enough market, there is no reason to penalize its customers with a software technology that is less effective than it could be, just because others will not be able to benefit from it.

Future Collaborations

One major goal of the seminar was to encourage participants of the different research communities to collaborate on stencil research in the future. Paper titles for a potential postproceedings volume were negotiated and collected in a discussion session on Thursday evening. Until its appearance, details remain confidential.

5 Conclusions

With 46 participants, the seminar was fully booked. A good spread of contemporary projects on stencil codes and high-performance DSLs was represented:

- from SPPEXA: ExaStencils, EXA-DUNE, TERRA-NEO, and ESSEX
- from elsewhere: Pochoir, PATUS, SPIRAL, STELLA, PolyMage, PyOP2, Polly

The main challenge was to reach an understanding between the members of the mutual needs of the diverse research communities – math, CS, applications. The culmination of the seminar’s success would consist of the appearance of a postproceedings and of sustained future collaborations.

Participants

- Sven Apel
Universität Passau, DE
- Mauro Bianco
CSCS – Lugano, CH
- Matthias Bolten
Bergische Univ. Wuppertal, DE
- Uday Bondhugula
Indian Institute of Science –
Bangalore, IN
- Rezaul Chowdhury
Stony Brook University, US
- Simplicio Donfack
University of Lugano, CH
- Christian Engwer
Universität Münster, DE
- Robert D. Falgout
LLNL – Livermore, US
- Franz Franchetti
Carnegie Mellon University –
Pittsburgh, US
- Michael Freitag
Universität Passau, DE
- Stephanie Friedhoff
KU Leuven, BE
- Francisco Jose Gaspar
University of Zaragoza, ES
- Björn Gmeiner
Univ. Erlangen-Nürnberg, DE
- Alexander Grebhahn
Universität Passau, DE
- Armin Größlinger
Universität Passau, DE
- Tobias Grosser
ETH Zürich, CH
- Gundolf Haase
Universität Graz, AT
- Georg Hager
Univ. Erlangen-Nürnberg, DE
- Frank Hannig
Univ. Erlangen-Nürnberg, DE
- Juraj Kardos
Brno Univ. of Technology, CZ
- Paul H. J. Kelly
Imperial College London, GB
- Hans-Peter Kersken
DLR – Köln, DE
- Harald Köstler
Univ. Erlangen-Nürnberg, DE
- Stefan Kronawitter
Universität Passau, DE
- Sebastian Kuckuk
Univ. Erlangen-Nürnberg, DE
- Bradley C. Kuszmaul
MIT – Cambridge, US
- Christian Lengauer
Universität Passau, DE
- Dmitry Mikushin
University of Lugano, CH
- Marcus Mohr
LMU München, DE
- Carlos Osuna Escamilla
ETH Zürich, CH
- Ekanathan Palamadai
Natarajan
MIT – Cambridge, US
- Dirk Pflüger
Universität Stuttgart, DE
- Louis-Noël Pouchet
Ohio State University –
Columbus, US
- Hannah Rittich
Bergische Univ. Wuppertal, DE
- Carmen Rodrigo Cardiel
University of Zaragoza, ES
- Ulrich Rüde
Univ. Erlangen-Nürnberg, DE
- P. Saday Sadayappan
Ohio State University –
Columbus, US
- Olaf Schenk
University of Lugano, CH
- Christian Schmitt
Univ. Erlangen-Nürnberg, DE
- Armando Solar-Lezama
MIT – Cambridge, US
- Jürgen Teich
Univ. Erlangen-Nürnberg, DE
- Wim Vanroose
University of Antwerp, BE
- Christian Waluga
TU München, DE
- Gabriel Wittum
Universität Frankfurt, DE
- Barbara Wohlmuth
TU München, DE
- Ulrike Meier Yang
LLNL – Livermore, US



Software and Systems Traceability for Safety-Critical Projects

Edited by

Jane Cleland-Huang¹, Sanjai Rayadurgam², Patrick Mäder³, and Wilhelm Schäfer⁴

1 DePaul University – Chicago, US, jhuang@cs.depaul.edu

2 University of Minnesota – Minneapolis, US, rsanjai@cs.umn.edu

3 TU Ilmenau, DE, patrick.maeder@tu-ilmenau.de

4 Universität Paderborn, DE, wilhelm@uni-paderborn.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15162 on “Software and Systems Traceability for Safety-Critical Projects”. The event brought together researchers and industrial practitioners working in the field of safety critical software to explore the needs, challenges, and solutions for Software and Systems Traceability in this domain. The goal was to explore the gap between the traceability prescribed by guidelines and that delivered by manufacturers, and starting from a clean slate, to clearly articulate traceability needs for safety-critical software systems, to identify challenges, explore solutions, and to propose a set of principles and domain-specific exemplars for achieving traceability in safety critical systems.

Seminar April 12–17, 2015 – <http://www.dagstuhl.de/15162>

1998 ACM Subject Classification D.2.4 Software/Program Verification

Keywords and phrases safety-critical software development, assurance cases, software and systems traceability

Digital Object Identifier 10.4230/DagRep.5.4.76

Edited in cooperation with Patrick Rempel

1 Executive Summary

Jane Cleland-Huang

Sanjai Rayadurgam

Patrick Mäder

Wilhelm Schäfer

License  Creative Commons BY 3.0 Unported license

© Jane Cleland-Huang, Sanjai Rayadurgam, Patrick Mäder, and Wilhelm Schäfer

Safety-critical systems, defined as systems whose “failure could result in loss of life, significant property damage, or damage to the environment”¹, pervade our society. Developing software is a challenging process. Not only must the software deliver the required features, but it must do so in a way that ensures that the system is safe and secure for its intended use. To this end safety-critical systems must meet stringent guidelines before they can be approved or certified for use. For example, software developed for the aerospace industry must comply to the

¹ *Failure Analysis and the Safety-Case Lifecycle*, W. S. Greenwell, E. A. Strunk, and J. C. Knight in *Human Error, Safety and Systems Development*, 2004, http://dx.doi.org/10.1007/1-4020-8153-7_11.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Software and Systems Traceability for Safety-Critical Projects, *Dagstuhl Reports*, Vol. 5, Issue 4, pp. 76–97

Editors: Jane Cleland-Huang, Sanjai Rayadurgam, Patrick Mäder, and Wilhelm Schäfer



DAGSTUHL
REPORTS Dagstuhl Reports

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

ISO12207 and/or the DO-178B/C guidelines, while software developed for European railway communication, signaling, and processing systems, must comply to EN50128. Most guidelines prescribe a set of steps and deliverable documents that focus around planning, analysis and design, implementation, verification and validation, configuration management, and quality assurance activities. In addition they often provide specific guidelines for the creation and use of traceability in the project. For example, depending upon the criticality level of a requirement, the US Federal Aviation Authority guideline DO-178B requires traceability from requirements to design, and from requirements to source code and executable object code.

In practice, traceability is achieved through the creation and use of *trace links*, defined by the Center of Excellence for Software and Systems Traceability² as “specified associations between pair of artifacts, one comprising the source artifact and one comprising the target artifact”. Software traceability serves an important role in demonstrating that a delivered software system satisfies its software design constraints and mitigates all identified hazards. When correct, traceability demonstrates that a rigorous software development process has been established and systematically followed. Current guidelines, in many safety-critical industries, prescribe traceability for two reasons. First, as an indirect measure that good practice has been followed, the general idea being that traceability information serves as an indicator that design and production practices were conducted in a sound fashion; and second, as a more direct measure, to show that specific hazards have been explored, potential failure modes identified, and that the system is designed and implemented in a “demonstrably rational way”.

Unfortunately, there is a significant gap between prescribed and actual traceability. An analysis of the traceability information submitted by various organizations to the US Food and Drug Administration (FDA) as part of the medical device approval process³, showed a significant *traceability gap* between the traceability expectations as laid out in the FDA’s “Guidance for the Content of Premarket Submissions for Software Contained in Medical Devices”, and the traceability data documented in the submissions. While all of the submissions made some attempt to satisfy the FDA’s traceability guidelines, serious deficiencies were found in almost all the submissions in terms of missing traceability paths, missing and redundant links, and problems in trace granularity. These deficiencies made it very difficult to understand the rationale for individual links. A more recent systematic analysis of seven software projects that originated from four different domains (automotive, aviation, medical, and space) revealed similar problems⁴. The provided software development artifacts were analyzed with respect to four technical guideline documents (ISO 26262-6, DO-178B, FDA Guide for Submissions, ECSS-E-40), where each document is a representative guideline of one of the four domains.

Problems are exacerbated in the systems engineering domain in which core concepts and designs are often documented across multiple models, each of which might depict a single viewpoint or perspective of the system. For example, the system might include separate models for functional and behavioral requirements, software components, electrical components, thermodynamics, and mechanical components. Furthermore, although each

² Center of Excellence for Software and Systems Traceability (<http://www.CoEST.org>)

³ *Strategic traceability for safety-critical projects*, P. Mäder, P. L. Jones, Y. Zhang, and J. Cleland-Huang, IEEE Software, 30(3):58–66, <http://dx.doi.org/10.1109/MS.2013.60>.

⁴ *Mind the gap: Assessing the conformance of software traceability to relevant guidelines*, P. Rempel, P. Mäder, T. Kuschke, and J. Cleland-Huang, Proc. of the 36th Int’l Conf. on Software Engineering (ICSE’14), <http://dx.doi.org/10.1145/2568225.2568290>.

of these perspectives is modeled separately in isolation from one another, they interact to produce the final behavior of the system. Traceability solutions must extend across these heterogeneous models. Deficiencies in traceability are certainly not new. As far back as 1995, Gotel et al. identified several different traceability problems and attributed them to poor coordination, lack of perceived benefits, time to market pressures, and lack of sufficient tooling. These problems observed almost 20 years ago, continue to plague the traceability landscape today, meaning that the *traceability gap* between what is prescribed and what is practiced is still very real.

Given that the software and systems engineering communities have been unable to solve this problem in over 20 years, it seems prudent to reexamine traceability needs and their prescribed solutions. Within this Dagstuhl seminar, we engaged software and systems engineering researchers and practitioners from the safety-critical domain alongside traceability experts, in highly focused discussions. The aim was to gain a deeper understanding of exactly what traceability is needed for safety-critical systems, and to identify practical and achievable solutions. To the best of our knowledge this was the first time researchers from the safety-critical and traceability domains came together in a dedicated forum to tackle this problem.

We started the week with a number of more general presentations and discussions from experts in the respective areas to form a common understanding for later discussions. Subsequently, the seminar continued with shorter talks focusing on a variety of specific aspects of open challenges and potential solutions accompanied by intensive and highly interactive discussions. In parallel, we parted for about one third of the time into four focus groups working on what had been identified as the most relevant and urgent challenges for closing the traceability gap. The four areas of focus were: tracing qualities, traceability in the context of models and tools, cost-benefit and stakeholder perspectives, and traceability in the context of evolution and change. In result, we intend to publish a white-paper that systematically analyzes the existing traceability gap based on the outcome of the four focus groups. Furthermore, the workshop has initiated collaborations and potential research projects between previously separate areas with the potential of significant impact.

2 Table of Contents

Executive Summary

Jane Cleland-Huang, Sanjai Rayadurgam, Patrick Mäder, and Wilhelm Schäfer . . . 76

Overview of Talks

Reusing Traceability for Change Impact Analysis – A Case Study in a Safety Context <i>Markus Borg</i>	81
Questioning the Traceability Requirements of Certifying Bodies <i>Jane Cleland-Huang</i>	81
Towards a Categorical Foundation of Model Synchronization <i>Krzysztof Czarnecki</i>	82
Model-to-Model Traceability as a Key Enabler for Domain-Specific Safety Analysis <i>Christopher Gerking</i>	84
Runtime Traceability Challenges in Systems of Systems <i>Paul Gruenbacher</i>	84
Traceability and the CoWolf framework <i>Lars Grunske</i>	84
Model-based Reliability and Safety Engineering <i>Kai Hoefig</i>	85
The Benefits of Traceability During Software Implementation <i>Patrick Maeder</i>	85
Model-based design inspection based on traceability information models and design slicing <i>Shiva Nejati</i>	87
Traceability Through Precise Process Definitions <i>Leon J. Osterweil</i>	87
Evolving Trace Links across Versions of a Software System in Safety-Critical Domain <i>Mona Rahimi</i>	88
Medical Device Verification and Validation: Experiences and Perspectives <i>Sanjai Rayadurgam</i>	88
Traceability Assessment and Roadmap for Medical Device Domain <i>Gilbert Regan</i>	89
Mind the Gap: Assessing the Conformance of Software Traceability to Relevant Guidelines <i>Patrick Rempel</i>	91
An Analysis of Challenges in Safety Certification and Implications for Traceability Research <i>Mehrdad Sabetzadeh</i>	92
Traceability in the Nuclear Energy Industry. Challenges and Lessons Learned from an Industrial Project <i>Nicolas Sannier</i>	92

Systems Engineering and Traceability at the Model Level <i>Wilhelm Schäfer</i>	94
Gene-Auto & QGen: Experiences and ideas on ACG specification, qualification and verification <i>Andres Toom</i>	94
Model-based safety engineering: Challenges and opportunities in practice <i>Marc Zeller</i>	96
Participants	97

3 Overview of Talks

3.1 Reusing Traceability for Change Impact Analysis – A Case Study in a Safety Context

Markus Borg (Lund University, SE)

License  Creative Commons BY 3.0 Unported license

© Markus Borg

Joint work of Borg, Markus; Wnuk, Krzysztof; Regnell, Björn; Runeson, Per

Change Impact Analysis (CIA) during software evolution of safety-critical systems is a fundamental task closely related to traceability. However, CIA is difficult and labor-intensive for complex systems, and several authors have proposed tool support. Unfortunately, very few have been evaluated in industrial settings. In this talk, I will introduce our tool ImpRec, a Recommendation System for Software Engineering (RSSE), tailored for CIA at an automation company. Building on research from assisted tracing using information retrieval solutions, as well as mining software repositories, ImpRec recommends development artifacts potentially impacted when resolving incoming issue reports. In contrast to previous work on automated CIA, our approach explicitly targets development artifacts that are not source code. I will present results from the evaluation of ImpRec, designed as a two-phase industrial case study. In the first part, we measured the correctness of ImpRec's recommendations by simulating the historical inflow of 12 years' worth of issue reports in the company. In the second part, we assessed the utility of working with ImpRec by deploying the RSSE in two development teams. Our results suggest that ImpRec presents about 40% of the true impact among the top-10 recommendations. Furthermore, user log analysis indicates that ImpRec can support CIA in industry, and the developers in our study also acknowledged the value of ImpRec in interviews. In conclusion, our findings show the potential of reusing traceability associated with developers' past activities in an RSSE. However, more research is needed on how to retrain the tool once deployed, and how to adapt processes when new tools are introduced in safety-critical contexts.

3.2 Questioning the Traceability Requirements of Certifying Bodies

Jane Cleland-Huang (DePaul University – Chicago, US)

License  Creative Commons BY 3.0 Unported license

© Jane Cleland-Huang

Software traceability is a sought-after, yet often elusive quality. It is required in safety-critical systems by many certifying and/or approving bodies, such as the USA Federal Aviation Authority (FAA) or the USA Food and Drug Administration (FDA). However, our previous study of medical device submissions to the FDA [1] highlighted the fact that adequate traceability was rarely achieved in practice. We identified numerous traceability problems including missing traceability paths, missing individual trace links, redundant paths, and inconsistencies. Furthermore, conversations with practitioners revealed that traceability is frequently built into a system as an afterthought, primarily for compliance purposes. In short, practitioners often perceive traceability effort as a burden, and rarely realize its benefits for supporting a broad range of software engineering activities for querying project data.

A traceability gap exists between what is prescribed by certifiers and what is delivered by product manufacturers [2]. This gap has several root causes. The traceability prescribed

by certifiers tends to be overly extensive, requiring traceability paths between almost every pair of artifact types but providing only weak rationales for each path. Practitioners often fail to understand the need for such extensive traceability and therefore deliver only a subset of the requested trace links. This tends to be accepted by certifiers thereby reinforcing the practice of delivering less than the prescribed traceability. Further, current traceability tools fall short of providing adequate support for trace link construction and maintenance. Trace links are created using drag-and-drop mechanisms while potentially outdated links are as ‘suspect’ whenever the source or target artifacts are modified. Significant manual effort is therefore needed to establish and maintain traceability. More promising state-of-the-art approaches capture trace links as a natural byproduct of the workflow, for example by requiring developers to tag work requests with the associated code. The high cost of traceability can also potentially be reduced using automated trace retrieval, or by utilizing more intelligent traceability solutions capable of integrating natural language processing techniques and domain knowledge in order to reason about the presence of links.

Traceability serves two primary purposes in safety-critical systems, to demonstrate that due process has been followed, and to assess whether specific hazards, regulatory codes, and/or mitigating requirements have been implemented in the design. However, it is not clear whether current certification guidelines capture ideal traceability requirements. Recent interest in building safety- and assurance-cases suggests that a better approach might focus traceability efforts on connecting hazards, claims, and evidence for those claims in order to demonstrate product safety while simultaneously showing that good process was followed.

Solving the traceability gap is going to require a multi-pronged effort. Tracing practices will need to become a natural byproduct of the software engineering process. State of the art solutions for retrieving and/or reconstructing missing trace links will need to improve so that the links they generate can be trusted by human users. Finally, certifying bodies will need to rethink their prescribed traceability requirements – so that any cost and effort involved in the traceability process returns clear benefits to both developers and certifiers.

References

- 1 J. Cleland-Huang, O. Gotel, J.H. Hayes, P. Mäder, and Zisman, A (2014). “Software Traceability: Trends and Future Directions”. In: *Proc. of FOSE’14/ICSE’14*, pp. 55–69.
- 2 P. Mäder, P.L. Jones, Y. Zhang, and J. Cleland-Huang (2013). “Strategic Traceability for Safety-Critical Projects”, In: *IEEE Software*, 30(3), pp. 58–66.

3.3 Towards a Categorical Foundation of Model Synchronization

Krzysztof Czarnecki (University of Waterloo, CA)

License  Creative Commons BY 3.0 Unported license
© Krzysztof Czarnecki

Main reference Z. Diskin, A. Wider, H. Gholizadeh, K. Czarnecki, “Towards a Rational Taxonomy for Increasingly Symmetric Model Synchronization,” in *Proc. of the 7th Int’l Conf. on Theory and Practice of Model Transformations (ICMT’14)*, pp. 57–73, Springer, 2014.

URL http://dx.doi.org/10.1007/978-3-319-08789-4_5

Model-driven engineering usually requires many overlapping models of a system, each supporting a particular kind of stakeholder or task. The consistency among these models needs to be managed during system development. Consistency management unfolds in the space of multiple model replicas, versions over time, different modeling languages, and complex relations among the models, which make the process complex and challenging.

This talk reports on our ongoing work to develop the theoretical foundation for model synchronization based concepts from category theory and illustrated its application to

practical model synchronization problems. This theory views model synchronization as an algebra, abstracting from any concrete model structures and synchronization function implementations and with laws regulating the properties of the functions.

We have delivered several pieces of such foundation, including a precise notion of model overlap for two or more models expressed in one or more languages [4], a general notion of model mapping based on queries [7], and a general framework of delta lenses [5, 6, 8]. A lens, originally proposed by Benjamin Pierce et al., is a coordinated pair of functions: get to extract an abstract view from a source artifact and put to update the source to make it consistent with an updated view. In contrast to the original state-based lenses, delta lenses operate on model updates (represented as vertical model mappings) and traces relating the overlapping models (represented as horizontal model mappings) rather than model states only. Basing lenses on model mappings addresses limitations of the state-based setting, such as composition anomalies and inflexible signatures of the propagation functions [5]. We have also addressed the symmetric case of delta lenses [6], solving a long-standing problem of bidirectional transformations: too strong PUTPUT/undoability laws. We have also shown that our delta lens framework can be instantiated in the Triple-Graph-Grammar (TGG) setting, giving necessary and sufficient correctness conditions, checkable by tools [8]. Based on these concepts, we have recently constructed a design space of model synchronizers, capturing fundamental design choices, such as incrementality, and informational and organizational symmetry [10].

Practical application of delta lenses include bi-directional synchronization between models and code [1, 2] and between models specifying business processes and executable models implementing the specifications [11].

References

- 1 M. Antkiewicz and K. Czarnecki (2006). “Framework-Specific Modeling Languages with Round-Trip Engineering”. In: *Proc. of MODELS’06*, 693–706.
- 2 M. Antkiewicz, T. Bartolomei, and K. Czarnecki (2009). “Fast extraction of high-quality framework-specific models from application code”. In: *ASE*, 16(1), 101–144.
- 3 M. Antkiewicz and K. Czarnecki (2008). “Design Space of Heterogeneous Synchronization”. In: *GTTSE’07, LNCS 5235*, 10, 3–46
- 4 Z. Diskin, Y. Xiong, and K. Czarnecki (2010). “Specifying Overlaps of Heterogeneous Models for Global Consistency Checking”. In: *Proc. of MDI Workshop*, 42–51
- 5 Z. Diskin, Y. Xiong, and K. Czarnecki (2011). “From State- to Delta-Based Bidirectional Model Transformations: the Asymmetric Case”. In: *JOT*, 10(6), 1–25
- 6 Z. Diskin, Y. Xiong, K. Czarnecki, H. Ehrig, F. Hermann, and F. Orejas (2011). “From State- to Delta-based Bidirectional Model Transformations: the Symmetric Case”. In *Proc. of MODELS’11*, 304–318
- 7 Z. Diskin, T. Maibaum, and K. Czarnecki (2012). “Intermodeling, queries, and Kleisli categories”. In: *Proc. of FASE’12*, 163–177
- 8 F. Hermann, H. Ehrig, F. Orejas, F. Czarnecki, Z. Diskin, Y. Xiong, S. Gottmann, and T. Engel (2013). “Model Synchronization Based on Triple Graph Grammars: Correctness, Completeness and Invertibility”. In: *SOSYM*, 14(1), 241–269
- 9 M. Branco, Y. Xiong, K. Czarnecki, J. Kuester, H. Voelzer (2014). “A case study on consistency management of business and IT process models in banking”. In: *SOSYM*, 13(3), 913–940
- 10 Z. Diskin, A. Wider, H. Gholizadeh, K. Czarnecki (2014). “Towards a Rational Taxonomy for Increasingly Symmetric Model Synchronization”. In: *Proc. of ICMT’14*, 57–73
- 11 J. Küster, H. Völzer, C. Favre, M. Branco, K. Czarnecki (2015). “Supporting Different Process Views through a Shared Process Model”. In: *SOSYM*, 25 pages

3.4 Model-to-Model Traceability as a Key Enabler for Domain-Specific Safety Analysis

Christopher Gerking (Universität Paderborn, DE)

License  Creative Commons BY 3.0 Unported license
© Christopher Gerking

Joint work of Gerking, Christopher; Dziwok, Stefan; Heinzemann, Christian; Schäfer, Wilhelm

Safety-critical systems raise the need for formal verification at an early stage of the design process. Model checking is a verification technique that provides counterexamples in case of violated safety properties. Domain-specific model checking (DSMC) [1] hides the complexity of model checking by translating from a domain-specific language (DSL) to the input of a given model checker, and using traceability information to translate counterexamples back to the DSL. Our approach addresses the problem that existing settings assume only minor differences between DSL and model checking language, which allows for a single-step translation. This talk demonstrates how model-to-model traceability enables a back-translation of counterexamples even in case of major differences between DSL and model checking language. Our case study describes a successful application of DSMC to a multi-step translation scenario from the domain of interconnected cyber-physical systems.

References

- 1 W. Visser, M. B. Dwyer, and M. W. Whalen (2012). “The hidden models of model checking”. In: *Software & Systems Modeling*, 11(4), 541–555

3.5 Runtime Traceability Challenges in Systems of Systems

Paul Gruenbacher (Universität Linz, AT)

License  Creative Commons BY 3.0 Unported license
© Paul Gruenbacher

This talk addresses challenges of using traceability links at runtime to diagnose problems in systems of systems (SoS). Specifically, it addresses traceability needs in setting up a requirements monitoring infrastructure for a system-of-systems architecture. The challenges are illustrated with examples from an industrial system of systems in the domain of metallurgical plants. Specifically, automated traceability techniques can support engineers defining requirements monitoring models. Better traceability between requirements and the SoS runtime architecture can further improve problem diagnoses after detecting violations of requirements.

3.6 Traceability and the CoWolf framework

Lars Grunske (Universität Stuttgart, DE)

License  Creative Commons BY 3.0 Unported license
© Lars Grunske

Agile and iterative development with changing requirements leads to continuously changing models. In particular, the researchers are faced with the problem of consistently co-evolving

different views of a model-based system. Whenever one model undergoes changes, corresponding models should co-evolve with respect to this change. On the other hand, domain engineers are faced with the huge challenge to find proper co-evolution rules which can be finally used to assist developers in the co-evolution process. In the presentation, the CoWolf framework is introduced that enables co-evolution actions between related models and provides a tooling environment. Furthermore, the results of a case study for the co-evolution of architecture and fault tree models [1] are presented.

References

- 1 S. Getir, A. Van Hoorn, L. Grunske, M. Tichy (2013). “Co-evolution of software architecture and fault tree models: An explorative case study on a pick and place factory automation system”. In: *Proc. of NiM-ALP@MoDELS'13*, 32–40

3.7 Model-based Reliability and Safety Engineering

Kai Hoefig (Siemens – München, DE)

License © Creative Commons BY 3.0 Unported license
© Kai Hoefig

Main reference K. Höfig, M. Zeller, L. Grunske, “metaFMEA-A Framework for Reusable FMEAs”, in *Proc. of the 4th Int'l Symp. on Model-Based Safety and Assessment (IMBSA'14)*, pp. 110–122, Springer, 2014.

URL http://dx.doi.org/10.1007/978-3-319-12214-4_9

Model driven development is currently one of the key approaches to cope with increasing development complexity, in general. Applying model-based approaches during the development of complex products aims at a systematic reuse of models or model elements and thus aims at a reuse of effort that already has been accomplished. A shorter time to market and decreased development costs are strong drivers from industry. Domain specific languages or model elements come into play to handle complexity and ease the development of systems. Domain specific and universal modeling languages provide purpose-oriented views on a system model. The ability to include variation points is used if product lines are being developed. The overall strategy of divide and conquer breaks complexity down into manageable parts.

Applying similar concepts to safety engineering is a promising approach to extend the advantages of model driven development to safety engineering activities aiming at a reduction of development costs, a higher product quality and a shorter time-to-market. First, it makes safety engineering as a standalone subtask of system development more efficient. Second, and even more important, this is an essential step towards a holistic development approach closing the gap between functional development and safety engineering.

3.8 The Benefits of Traceability During Software Implementation

Patrick Maeder (TU Ilmenau, DE)

License © Creative Commons BY 3.0 Unported license
© Patrick Maeder

Joint work of Maeder, Patrick; Egyed, Alexander

Main reference P. Mäder, A. Egyed, “Do developers benefit from requirements traceability when evolving and maintaining a software system?”, *Empirical Software Engineering*, 20(2):413–441, 2015.

URL <http://dx.doi.org/10.1007/s10664-014-9314-z>

Software traceability is a required component of many software development processes. Advocates of software traceability cite advantages like easier program comprehension and

support for software maintenance (i.e., software change). However, despite its growing popularity, for a long time there existed no published evaluation about the usefulness of requirements traceability. It is important, if not crucial, to investigate whether the use of requirements traceability can significantly support development tasks to eventually justify its costs [3, 1, 4, 5]. We thus conducted a controlled experiment with 71 subjects re-performing real implementation tasks on two third-party development projects: half of the tasks with and the other half without traceability. Our findings show that subjects with traceability performed on average 24% faster on a given task and created on average 50% more correct solutions [2, 6] – suggesting that traceability not only saves effort but can profoundly improve software implementation quality. For a follow-up study [7], we selected medium to large-scale open-source projects and focused especially on the discovered effect for implementation quality. We quantified for each developed component of each software project, the degree to which a studied development activity was enabled by existing traceability and set this metric in relation to the number of defects that occurred in a component. We found that traceability significantly affects the defect rate in a component. Overall, our results provide for the first time empirical evidence that traceability significantly improves implementation speed as well as implementation quality during software development.

References

- 1 P. Mäder, O. Gotel, and I. Philippow (2009). “Motivation matters in the traceability trenches”, In: *Proc. of RE’09*, pp. 143–148.
- 2 P. Mäder and A. Egyed (2012). “Assessing the effect of requirements traceability for software maintenance”, In: *Proc. of ICSM’12*, pp. 171–180.
- 3 E. Bouillon, P. Mäder, I. Philippow (2013). “A Survey on Usage Scenarios for Requirements Traceability in Practice”, In: *Proc. of REFSQ’13*, pp. 158–173.
- 4 P. Rempel, P. Mäder, and T. Kuschke (2013). “An empirical study on project-specific traceability strategies”, In: *Proc. of RE’13*, 195–204.
- 5 P. Rempel, P. Mäder, T. Kuschke, and I. Philippow (2013). “Requirements Traceability across Organizational Boundaries – A Survey and Taxonomy”, In: *Proc. of REFSQ’13*, pp. 125–140.
- 6 P. Mäder and A. Egyed (2015). “Do developers benefit from requirements traceability when evolving and maintaining a software system?”, In: *Empirical Software Engineering*, 20(2), pp. 413–441.
- 7 P. Rempel and P. Mäder (2015). “Estimating the Implementation Risk of Requirements in Agile Software Development Projects with Traceability Metrics”, In: *Proc. of REFSQ’15*, pp. 81–97.

3.9 Model-based design inspection based on traceability information models and design slicing

Shiva Nejati (University of Luxembourg, LU)

License © Creative Commons BY 3.0 Unported license
© Shiva Nejati

Joint work of Nejati, Shiva; Sabetzadeh, Mehrdad; Briand, Lionel; Falessi, Davide

Main reference S. Nejati, M. Sabetzadeh, D. Falessi, L. C. Briand, and T. Coq (2012). “A SysML-based approach to traceability management and design slicing in support of safety certification: Framework, tool support, and case studies”, *Information & Software Technology*, 54(6): 569–590, 2012.

URL <http://dx.doi.org/10.1016/j.infsof.2012.01.005>

Traceability is one of the basic tenets of all software safety standards and a key prerequisite for certification of software. Despite this, the safety-critical software industry is still suffering from a chronic lack of guidelines on traceability. An acute traceability problem that we have identified through observing the software safety certification process has to do with the link between safety requirements and software design. In the current state of practice, this link often lacks sufficient detail to support the systematic inspections conducted by the certifiers of the software safety documentation. As a result, the suppliers often have to remedy the traceability gaps after the fact which can be very expensive and the outcome might be far from satisfactory.

The goal of our work is to provide a traceability methodology and a design slicing algorithm for software safety certification by applying and specializing the Systems Modeling Language (SysML). Our methodology enables the establishment of traceability links prescribed by a traceability information model as well as a mechanism for extracting a minimized and relevant slice of the design with respect to the specified traceability links. The certifiers can then utilize the links and the design slices to effectively investigate safety claims. To validate our approach, we report on an industrial case study applying the approach to a safety IO software module used on ships and offshore facilities.

In this talk, I describe the context in which the above work was carried out, explain our proposed solutions, and discuss how our solutions have been applied to case studies from the Maritime and Energy domain.

3.10 Traceability Through Precise Process Definitions

Leon J. Osterweil (University of Massachusetts – Amherst, US)

License © Creative Commons BY 3.0 Unported license
© Leon J. Osterweil

Traceability should be viewed as a property needed support tracing, whose purpose should be viewed as the gathering and maintenance of key knowledge and understandings about software products. Precise and detailed definitions of the processes by which these products are developed, tested, and evolved are excellent vehicles for continuous creation and maintenance of the inter- and intra- artifact links that provide the desired traceability. This talk describes the Little-JIL process definition language and how it can be used to create processes that can be used to create these links and support this kind of traceability.

3.11 Evolving Trace Links across Versions of a Software System in Safety-Critical Domain

Mona Rahimi (DePaul University – Chicago, US)

License  Creative Commons BY 3.0 Unported license
© Mona Rahimi

Trace links provide critical support for numerous software engineering activities including safety analysis, compliance verification, test-case selection, and impact prediction in safety-critical systems. However, as the system evolves over time, there is a tendency for the quality of trace links to degrade into a tangle of inaccurate and untrusted links. This is especially true with the links between source-code and upstream artifacts such as requirements—because developers frequently refactor and change code without updating links. We present TLE(Trace Link Evolver), a solution for automating the evolution of trace links as the change is introduced to source code. We use a set of heuristics, open source tools and information retrieval methods to detect common change scenarios in different versions of code. Each change scenario is associated with a set of link evolution heuristics which are used to evolve trace links. We evaluated our approach through a controlled experiment and also through applying it to a selection of classes taken from Cassandra Database System. Results show that the trace links produced by our approach is significantly more accurate than those produced using information retrieval alone.

3.12 Medical Device Verification and Validation: Experiences and Perspectives

Sanjai Rayadurgam (University of Minnesota – Minneapolis, US)

License  Creative Commons BY 3.0 Unported license
© Sanjai Rayadurgam

Medical devices such as infusion pumps and pacemakers are safety-critical systems that are strictly regulated by government agencies. The safety of these devices must be demonstrably established prior to gaining approval for sale. Assurance cases, which are structured arguments that use evidence gathered through the course of development to establish desirable claims, are being used, and in some cases mandated, to establish safety of these devices. The feasibility and the merits of such arguments are critically dependent on traceability across all development process artifacts, which provide the evidence for the assurance case. Maintaining and evolving traceability information throughout the development process is challenging. In particular, showing that requirements are realized in specific design elements and that realization has been verified is necessary. Often, requirements and architecture co-evolve [4] and so attempting to specify one without the other leads to inconsistent specifications.

When the architecture is formally modeled and the requirements are decomposed along architectural lines, compositional verification techniques can be used to prove that the components satisfying their requirements and interacting as specified by the architecture, are sufficient to ensure that the system meets its requirements [3]. However, this is typically insufficient to make a complete argument for claiming verification. Components at the lowest levels of the architectural decomposition, have to be elaborated below into realizable implementations and their behavior verified or tested to check conformance to their respective requirements. Above the architectural model, there may be models of usage scenarios for

the system in its context, which may then be checked to validate that the system as specified would meet its intended needs. To tie these together into a complete satisfaction argument for safety claims [2], trace links that span multiple models and relate elements of that model at finer levels of granularity are needed. Further, the assurance arguments and consequently the trace links that enables argumentation must evolve along with the corresponding artifacts throughout development. In general, the models developed during the course of system building support several activities such as simulations, analysis, verification and code-generation. The relationships between the models, and the relationships established by these activities are essential to the assurance arguments. When architectural designs are annotated with logical rules of argumentation, generation of assurance cases can be automated [1].

A few observations related to traceability in this regard merit consideration. First, requirements placed on traceability solutions for safety-critical medical devices include both a stable core (that is mandated by regulations) as well as an evolving frontier (that is driven by changes in development methods and tools employed). Second, while automation can speed up several traceability related tasks, it is not helpful especially when the output produced has to be manually analyzed. Strategic thinking is needed to strike a good balance between automation and manual analysis. Third, questions of provenance tend to be difficult to answer, but those are the most useful for constructing good assurance arguments. Fourth, supporting multiple viewpoints of various stakeholders during development requires incremental evolution of trace semantics because the stakeholder requirements also evolve throughout development.

References

- 1 A. Gacek, J. Backes, D. Cofer, K. Slind, and M. Whalen (2014). “Resolute: An assurance case language for architecture models”, In: *Proc. of HILT '14*, pp. 19–28.
- 2 A. Murugesan, O. Sokolsky, S. Rayadurgam, M. Whalen, M. Heimdahl, and I. Lee (2014). “Linking abstract analysis to concrete design: A hierarchical approach to verify medical CPS safety”, In: *Proc. of ICCPS'14*, pp. 139–150.
- 3 A. Murugesan, M. W. Whalen, S. Rayadurgam, and M. Heimdahl (2013). “Compositional verification of a medical device system”, In *Ada Lett.*, 33(3), pp. 51–64.
- 4 M.W. Whalen, A. Gacek, D. Cofer, A. Murugesan, M. Heimdahl, and S. Rayadurgam (2013). “Your ‘What’ Is My ‘How’: Iteration and Hierarchy in System Design”, In: *IEEE Software*, 30(2), pp. 54–60.

3.13 Traceability Assessment and Roadmap for Medical Device Domain

Gilbert Regan (Dundalk Institute of Technology, IE)

License © Creative Commons BY 3.0 Unported license
© Gilbert Regan

Within the medical device domain, as in other safety critical domains, software must provide reliability, safety and security because failure to do so can lead to injury or death. Software is a complex element of a medical device; it’s role, functionality and importance continually increases. Additionally due to changes in the 2007 medical device directive (MDD 2007/47/EC), standalone software can now be classed as an active medical device in its own right. Developing medical device software-based systems in a disciplined and cost-effective way poses major challenges (esp. with the move towards mobile devices, patient-driven applications, wireless devices and cloud-based solutions). Therefore highly effective software

practices are required. Additionally, regulation normally requires safety critical systems are certified before entering service. This involves submission of a safety case (a reasoned argument that the system is acceptably safe) to the regulator. A safety case should include evidence that the organization has established effective software development processes that are based on recognized engineering principles appropriate for safety critical systems. At the heart of such processes, they must incorporate traceability.

However, numerous barriers hamper the effective implementation of traceability such as cost, complexity of relationship between artifacts, calculating a return on investment, different stakeholder viewpoints, lack of awareness of traceability and a lack of guidance on what traceability to implement and how to implement it. There are a number of standards which medical device manufacturers must conform to, however these standards have different traceability requirements. This leads to confusion as to what traceability manufacturers should implement. Additionally medical device software manufacturers are often very small organisations with little experience in traceability, and, in addition to ‘what traceability to implement’, they are often unsure as to ‘how to implement it’.

In Ireland the importance of the medical device Industry is obvious from its contribution of 8.5% of Ireland’s total merchandise exports, and this sector has been identified as one of the key drivers of industrial growth for the future. Consequently the Irish government fund research into how medical device organizations can improve their software development process. The implementation of traceability through the software development lifecycle and supporting processes of risk management and change management has been identified as a weakness within these medical device organizations. To assist these organizations improve their implementation of traceability, a decision was taken to address the ‘lack of guidance’ on what traceability to implement and how to implement it. This decision lead to the development of the following research question: “To what extent can the development of a traceability assessment and implementation framework assist medical device software organizations improve their traceability practices and put them on the path to regulatory compliance?”

To answer this question a traceability process assessment model and a roadmap for the implementation of traceability have been developed. In this presentation the experience of developing and trialling a traceability assessment model in two medical device organizations is presented. We show that the assessment model was successful in identifying strengths and weaknesses in both organisations implementation of traceability. Additionally a roadmap to assist organizations implement traceability that is both efficient and compliant is presented. Finally, through the experience of trialling the assessment model in two medical device organizations, I think the traceability assessment model could be improved through automation and so propose an initial idea of using the Open Services for Lifecycle Collaboration (OSLC) initiative.

3.14 Mind the Gap: Assessing the Conformance of Software Traceability to Relevant Guidelines

Patrick Rempel (TU Ilmenau, DE)

License © Creative Commons BY 3.0 Unported license
© Patrick Rempel

Joint work of Rempel, Patrick; Mäder, Patrick; Kuschke, Tobias; Cleland-Huang, Jane

Main reference P. Rempel, P. Mäder, T. Kuschke, J. Cleland-Huang, “Mind the Gap: Assessing the Conformance of Software Traceability to Relevant Guidelines”, in *Proc. of the 36th Int’l Conf. on Software Engineering (ICSE’14)*, pp. 943–954, ACM, 2014.

URL <http://dx.doi.org/10.1145/2568225.2568290>

Many guidelines for safety-critical industries such as aeronautics, medical devices, and railway communications, specify that traceability must be used to demonstrate that a rigorous process has been followed and to provide evidence that the system is safe for use. However, practitioners rarely follow explicit traceability strategies [2, 1]. Organizations struggle to establish and maintain accurate and complete sets of traceability links [3, 4]. In practice, there is a gap between what is prescribed by guidelines and what is implemented in practice, making it difficult for organizations and certifiers to fully evaluate the safety of the software system [5]. We present an approach, which parses a guideline to extract a Traceability Model depicting software artifact types and their prescribed traces. It then analyzes the traceability data within a project to identify areas of traceability failure [7]. Missing traceability paths, redundant and/or inconsistent data, and other problems are highlighted. We used our approach to evaluate the traceability of seven safety-critical software systems and found that none of the evaluated projects contained traceability that fully conformed to its relevant guidelines [6].

References

- 1 P. Mäder, O. Gotel, and I. Philippow (2009). “Getting back to basics: Promoting the use of a traceability information model in practice”, In: *Proc. of TEFSE@ICSE’09*, pp. 143–148.
- 2 P. Mäder, O. Gotel, and I. Philippow (2009). “Motivation matters in the traceability trenches”, In: *Proc. of RE’09*, pp. 143–148.
- 3 P. Rempel, P. Mäder, and T. Kuschke (2013). “An empirical study on project-specific traceability strategies”, In: *Proc. of RE’13*, pp. 195–204.
- 4 P. Rempel, P. Mäder, T. Kuschke, and I. Philippow (2013). “Requirements Traceability across Organizational Boundaries – A Survey and Taxonomy”, In: *Proc. of REFSQ’13*, pp. 125–140.
- 5 P. Mäder, P. L. Jones, Y. Zhang, and J. Cleland-Huang (2013). “Strategic Traceability for Safety-Critical Projects”, In: *IEEE Software*, 30(3), pp. 58–66.
- 6 P. Rempel, P. Mäder, T. Kuschke, and J. Cleland-Huang (2014). “Mind the Gap: Assessing the Conformance of Software Traceability to Relevant Guidelines”, In: *Proc. of ICSE’14*, pp. 943–954.
- 7 P. Rempel and P. Mäder (2015). “A Quality Model for the Systematic Assessment of Requirements Traceability”, In: *Proc. of RE’15*, 8 pages.

3.15 An Analysis of Challenges in Safety Certification and Implications for Traceability Research

Mehrdad Sabetzadeh (University of Luxembourg, LU)

License © Creative Commons BY 3.0 Unported license
© Mehrdad Sabetzadeh

Many safety-critical systems in domains such as healthcare, aviation, and railways are subject to safety certification. The goal of safety certification is to provide confidence that a system will function safely in the presence of known hazards. Safety certification can be associated with the assessment of products, processes, or personnel. For software-intensive safety-critical systems, certification of products and processes are the most challenging.

In my talk, I will discuss several challenges in the safety certification of software-intensive systems. These challenges were gleaned from a large systematic review of the academic literature on safety certification, a number of practitioner surveys, and my personal experience working with the safety-critical software industry. I will argue that safety certification is closely intertwined with traceability, and that many of the challenges faced in safety certification today are caused by traceability gaps.

At the end, I will briefly present some technical work from our research group lying at the intersection of safety certification and traceability research.

3.16 Traceability in the Nuclear Energy Industry. Challenges and Lessons Learned from an Industrial Project

Nicolas Sannier (University of Luxembourg, LU)

License © Creative Commons BY 3.0 Unported license
© Nicolas Sannier

Joint work of Sannier, Nicolas; Baudry, Benoit
Main reference N. Sannier, B. Baudry, “INCREMENT: A Mixed MDE-IR Approach for Regulatory Requirements Modeling and Analysis,” in Proc. of the 20th Int’l Working Conf. on Requirements Engineering: Foundation for Software Quality (REFSQ’14), pp. 135–151, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-05843-6_11

The basic intuition behind any thermal power plant (independently of the primary resource they use: coal, gas, oil or uranium) is rather simple. Legitimate safety issues and safety measures to prevent catastrophic accidents or mitigate their consequences make these systems incredibly complex, and the more complex, the harder the safety qualification. Instrumentation and Control (I&C) systems allow to measure and control the plant’s behavior. Since 1986, I&C Systems are now mostly made of software systems and represent the toughest part for safety qualification. Those important to safety must conform to safety and regulatory requirements. Regulatory requirements are written by national safety authorities and are completed using a set of national recommendation guides or national and international standards. All these documents are weakly interrelated. Due to the lack of international consensus on regulatory practices, building such systems in different countries requires facing practices of several safety authorities. In order to minimize design and qualification effort, traceability between regulatory requirements became suddenly important. These observations set three important challenges. First, the global domain knowledge is scattered, not formalized and held by few experts. Second, traceability links and, said differently, the organization within the domain, is implicit. The third problem is the consequence of the two firsts. Bridges between different national practices are not developed, whereas the

understanding of regulations and practices becomes a significant industrial issue. In our context, traceability comes with two facets. Traceability basically means linking one element to another. However, two elements are basically linked for a particular purpose. In this case, traceability also means organizing the domain. (1) We need to define different types of traceability links between elements of our domain. It is one thing to define traceability links. It is another one to concretely present the relationship(s) between two artifacts as traceability links. In the modeling community, people are used with diagrams, where traceability between elements can be visualized as arrows between two boxes. However, for this particular project, nuclear engineers are mainly working with a large amount (thousands) of textual fragments, using excel spreadsheets. This representation put the text as first class citizen, and traceability cannot be handled properly in a class diagram neither in a matrix for such a huge amount of data. (2) We need to provide a convenient way to present these traceability links. Another concern worth considering is how to implement these traceability links. Said differently, it will be more than helpful to automatically build these traceability links. In our context, building traceability links between requirements means investigating relationships between requirements from the same corpus, and from different corpora in order to find similarities and possible coverage between requirements. In practice, these traceability links rarely exist in the regulation (regulations may have been written before the standards and were not always updated, practices may have changed). Moreover, these texts are generic, do not target any particular system and are voluntary ambiguous in order to last along the years. Finally regulations, standards and their interpretation as well as practices widely differ from one country to another. Consequently, there is no straightforward mapping between safety requirements. (3) Before creating traceability links between elements, we must first reduce the amount of elements to compare and analyze. In this project, we investigated different areas of research. We used Metamodeling to model the domain and precisely define the traceability links we wanted to represent. Modeling is good for formally defining domain elements and their relationship, however, it is not useful for representing information that is mainly textual and “would never fit in a box”, neither it is to analyze these textual fragments. For the latter, we investigated the use of different techniques such as overlapping clustering algorithms, machine learning and topic detection, and information retrieval. The objective was to be able to build topic clusters and reduce the size of the search space. We draw some observations from these experiments. In particular, machine learning and clustering approaches were considered suspicious by our industry partners, as defined topics could not be legitimately argued and validated. Results were “not good enough” to impose confidence. On the other hand, Information Retrieval performed reasonably well and received more positive feedbacks as the querying tool we proposed made more sense to them as it was in line with the support our industry partners were expecting. Among the many lessons that can be learned with respect to traceability, we can highlight the following ones:

- Traceability is not only a matter of linking objects together; it is also a matter of amount of objects to link together.
- Traceability techniques may also require trust and understandability to be accepted.

3.17 Systems Engineering and Traceability at the Model Level

Wilhelm Schäfer (Universität Paderborn, DE)

License  Creative Commons BY 3.0 Unported license
© Wilhelm Schäfer

Today's embedded and often safety-critical systems require traceability from requirements down to the implementation in terms of software and hardware. The talk presents a systematic V-model based approach which includes a discipline spanning model in the requirements and early design phase. This model consists of seven views which cover all disciplines as for example a so-called active structure. The active structure includes the definition of system components and energy, material and information flow between the components. Other views define scenarios or use cases and the underlying abstract shape model which comes from CAD. Partially automatic transformations which are based on a formal, semantically well-defined mechanism, define how discipline-specific models are derived. These transformations form the basis for defining traces between all concerned models.

3.18 Gene-Auto & QGen: Experiences and ideas on ACG specification, qualification and verification

Andres Toom (IB Krates OÜ – Tallinn, EE)

License  Creative Commons BY 3.0 Unported license
© Andres Toom

Automatic Code Generators (ACG) or model transformers make Model Driven Engineering (MDE) really powerful and have a great potential for reducing human errors and assisting certification by having an ability to generate together with the expected output also complete and consistent trace data. However, ACGs are complex tools themselves and need to be also qualified/certified.

This presentation reports on the experiences of two consecutive collaborative initiatives Gene-Auto [1] and its continuation carried out in two affiliated projects Project P [2] and Hi-MoCo [3], aiming at developing open-source code generators for safety critical domains. These code generators are designed to transform high-level modelling languages such as Simulink, Scicos and Stateflow to low-level program code in languages such as C or Ada. Since the intended end-user domains include avionics, their qualification plans have been set up according to the DO-178 B/C software qualification guideline. As this guideline is also one of the most stringent and refined ones among the safety critical domains, it is expected to be relatively easy to apply the results also in other domains. The outcomes of these initiatives are due to the effort of several organisations and many individuals over several years. References of the contributors can be found from the provided websites.

Gene-Auto was the first project, with most of the development taking place during an ITEA project in 2006-2008. The goals of the project included the clarification of the requirements for such a tool across different domains, investigating the qualification of Java language based software, and usage of formal methodologies such as development with a formal proof assistant in a qualifiable tool development process. Below is a brief summary of all of these outcomes.

The project went through several iterations and ended with a rather mature prototype ACG. The high-level user requirements and low-level software requirements (architecture and

component requirements) were considerably refined, but the global functional requirements (resulting from the toolset integration) were not yet explicitly specified. Similarly, source code and test cases were not explicitly traced to the requirements. For this project, differently from some more typical critical software development tasks, this was a rather natural process. As the requirements were largely clarified during an iterative process, rigorous management of trace links at all artifact levels was considered not that useful at that stage. Also, since the toolset had a very clear architectural design the risk of software deviating from high-level requirements was minimal.

The part of requirements that handled language definitions were formalised as metamodels and grammars. Other parts were specified as tagged Open Office documents. These documents were parsed and analysed by the Tramway (Topcased-Requirements) tool. The requirement split-down and coverage analysis performed by Tramway was rather useful. However, document based requirement management became soon complicated in terms of version and change management.

As for the qualification of Java-based software, then the main open questions remained related to the qualification of the Java Virtual Machine (JVM) and libraries, both native and external (such as parser generator, XML library). The advantage of Java is that it has a large user community and rich functionality. However, the libraries have much more functionality than is needed for a tool with concise, but safety critical functionality. One would either have to qualify or remove this extra functionality. Both options have considerable cost.

A specific procedure in the qualification plan was also set up for component development with a proof assistant. In particular the Coq proof assistant was used. In this process the initial component requirements specification phase is followed by a formal specification phase and a rather minimal design phase. Software code (with the exception of some interface and glue code) is automatically generated from the proof of the properties expressed in the formal specification. The program extraction technology has been developed out earlier and has been used on some tools of considerable complexity (e.g. the CompCert compiler by X. Leroy et al.). The certification experts considered the process outlined above acceptable for the qualification of such a tool possible in the long term. However, all the components used in the process, including the Coq kernel and program extractor need to be qualified. On the other hand, usage of such deep formal methods and tools in the industry was not considered possible by the Gene-Auto industrial partners at the current time.

The Gene-Auto initiative was continued in the joint follow-up projects Project P and Hi-MoCo, which laid the foundations for the QGen [4] code generator. The requirements of Gene-Auto have been refined and extended and different technical platforms are used to ease the tool qualification process. The main implementation language of the tool is Ada, which has been developed specifically for safety critical domains. Only a minimal set of already qualified external Ada libraries are used. Currently, the complete tool-chain has been implemented in Ada. The toolset has a standard Ecore metamodel-based interface for exporting-importing models between the transformation steps. This enables substituting elementary transformation steps by components implemented using other technologies (e.g. formal methods-based), when they are at a sufficient maturity level for tool qualification.

All qualification artifacts for QGen are managed using the Qualifying Machine (QM) [5] tool. The main functionalities of this tool currently include importing artifacts from different formats, analysing and displaying them, and (to some extent) also modifying via a common user interface. The low level requirements are written as structured annotations or formal pre/post conditions to the Ada spec (.ads) files that specify the public interface of source code modules of QGen. This way it is easy to ensure that all the public functions have

associated requirements, as well as update either the source code or low level requirement each time one of them changes. When requirements are expressed as formal invariants or pre-post conditions, mismatches between the requirement and implementation can be also automatically detected. Mapping between the test cases and requirements is achieved by explicit QM cross-links. Overall, at this stage the QGen tool maturation as well as qualification process and data refinement are still in progress. In addition we are performing complementary studies with more formal, but light-weight approaches that are close to the current state of the art and practice in the industry. These experiments include the formal specification of block libraries for dataflow languages [6] and transformation contracts for the specification of model transformations [7]. However, it is evident that most factors that complicated and impeded the qualification process of the Gene-Auto project have been refined and resolved in the QGen project.

References

- 1 Gene-Auto project (2006-2008). <http://www.geneauto.org>.
- 2 Project P (2011–2015). <http://www.open-do.org/projects/p>.
- 3 High-Integrity Model Compiler (Hi-MoCo) project (2011-2014). <http://www.eurekanetwork.org/project/-/id/6037>.
- 4 QGen tool (2015). <http://www.adacore.com/qgen>.
- 5 Qualifying Machine (QM) project (2015). <http://www.open-do.org/projects/qualifying-machine>.
- 6 A. Dieumegard, A. Toom, and M. Pantel (2014). “A software product line approach for semantic specification of block libraries in dataflow languages”, In: *Proc. of SPLC’14*, pp. 217–226.
- 7 A. Toom, A. Dieumegard, M. Pantel (2014). “Specifying and verifying model transformations for certified systems using transformation models”, In: *Proc. of ERTS2’14*.

3.19 Model-based safety engineering: Challenges and opportunities in practice

Marc Zeller (Siemens – München, DE)

License  Creative Commons BY 3.0 Unported license
© Marc Zeller

Joint work of Zeller, Marc; Hoefig, Kai

The technology path MbrSE develops and integrates models and methods for the model-driven engineering of critical systems. MbrSE stands for Model-based Reliability and Safety engineering and is primarily motivated by the challenges of dynamic reconfigurable cyber-physical systems.

We provide Siemens business units with top-notch technologies to establish systematic reuse of critical development artifacts and security-aware runtime certification.

Our methods and models enable divide and conquer strategies for critical systems development to reduce effort and increase quality of Siemens products.

Participants

- Markus Borg
Lund University, SE
- Jane Cleland-Huang
DePaul University – Chicago, US
- Krzysztof Czarnecki
University of Waterloo, CA
- Christopher Gerking
Universität Paderborn, DE
- Paul Grünbacher
Universität Linz, AT
- Lars Grunske
Universität Stuttgart, DE
- Kai Höfig
Siemens – München, DE
- Patrick Mäder
TU Ilmenau, DE
- Shiva Nejati
University of Luxembourg, LU
- Leon J. Osterweil
University of Massachusetts –
Amherst, US
- Mona Rahimi
DePaul University – Chicago, US
- Sanjai Rayadurgam
University of Minnesota –
Minneapolis, US
- Gilbert Regan
Dundalk Institute of Technology,
IE
- Patrick Rempel
TU Ilmenau, DE
- Mehrdad Sabetzadeh
University of Luxembourg, LU
- Nicolas Sannier
University of Luxembourg, LU
- Wilhelm Schäfer
Universität Paderborn, DE
- Andres Toom
IB Krates OÜ – Tallinn, EE
- Marc Zeller
Siemens – München, DE



Theory and Practice of SAT Solving

Edited by

Armin Biere¹, Vijay Ganesh², Martin Grohe³, Jakob Nordström⁴,
and Ryan Williams⁵

- 1 Universität Linz, AT, biere@jku.at
- 2 University of Waterloo, CA, vganesh@uwaterloo.ca
- 3 RWTH Aachen, DE, grohe@informatik.rwth-aachen.de
- 4 KTH Royal Institute of Technology, SE, jakobn@kth.se
- 5 Stanford University, US, rrwilliams@gmail.com

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15171 “Theory and Practice of SAT Solving”. The purpose of this seminar was to explore one of the most significant problems in all of computer science, namely that of computing whether formulas in propositional logic are satisfiable or not. This problem is believed to be intractable in general (by the theory of *NP*-completeness). However, the last two decades have seen dramatic developments in algorithmic techniques, and today so-called SAT solvers are routinely and successfully used to solve large-scale real-world instances in a wide range of application areas.

A surprising aspect of this development is that the best current SAT solvers are still to a large extent based on methods from the early 1960s, which can often handle formulas with millions of variables but may also get hopelessly stuck on formulas with just a few hundred variables. The fundamental question of when SAT solvers perform well or badly, and what underlying mathematical properties of the formulas influence SAT solver performance, remains very poorly understood. Another intriguing aspect is that much stronger mathematical methods of reasoning about propositional logic formulas are known today, in particular methods based on algebra and geometry, and these methods would seem to have great potential based on theoretical studies. However, attempts at harnessing the power of such methods have conspicuously failed to deliver any significant improvements in practical performance.

This seminar gathered leading researchers in applied and theoretical areas of SAT and computational complexity to stimulate an increased exchange of ideas between these two communities. We see great opportunities for fruitful interplay between theoretical and applied research in this area, and believe that this seminar showed beyond doubt that a more vigorous interaction between the two has potential for major long-term impact in computer science, as well for applications in industry.

Seminar April 19–24, 2015 – <http://www.dagstuhl.de/15171>

1998 ACM Subject Classification I.2.3 Deduction and Theorem Proving, F.2.2 Nonnumerical Algorithms and Problems: Complexity of proof procedures, F.4.1 Mathematical Logic, D.2.4 Software/Program Verification: Formal methods

Keywords and phrases SAT, Boolean SAT solvers, SAT solving, conflict-driven clause learning, Gröbner bases, pseudo-Boolean solvers, proof complexity, computational complexity, parameterized complexity

Digital Object Identifier 10.4230/DagRep.5.4.98

Edited in cooperation with Marc Vinyals (KTH Royal Institute of Technology)



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Theory and Practice of SAT Solving, *Dagstuhl Reports*, Vol. 5, Issue 4, pp. 98–122

Editors: Armin Biere, Vijay Ganesh, Martin Grohe, Jakob Nordström, and Ryan Williams



DAGSTUHL
REPORTS

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

1 Executive Summary

Armin Biere

Vijay Ganesh

Martin Grohe

Jakob Nordström

Ryan Williams

License © Creative Commons BY 3.0 Unported license
© Armin Biere, Vijay Ganesh, Martin Grohe, Jakob Nordström, and Ryan Williams

This seminar brought together researchers working in the areas of applied SAT solving on the one hand, and in proof complexity and neighbouring areas of computational complexity theory on the other, in order to communicate new ideas, techniques, and analysis from both the practical and theoretical sides.

The goals of this endeavour are to better understand why modern SAT solvers work so efficiently for many large-scale real-world instances, and in the longer term to discover new strategies for SAT solving that could go beyond the present “conflict-driven clause-learning” paradigm and deliver substantial further gains in practical performance.

Topics of the Workshop

This seminar explored one of the most significant problems in all of mathematics and computer science, namely that of proving logic formulas. This is a problem of immense importance both theoretically and practically. On the one hand, it is believed to be intractable in general, and deciding whether this is so is one of the famous million dollar Clay Millennium Problems (the P vs. NP problem). On the other hand, today so-called SAT solvers are routinely and successfully used to solve large-scale real-world instances in a wide range of application areas (such as hardware and software verification, electronic design automation, artificial intelligence research, cryptography, bioinformatics, operations research, and railway signalling systems, just to name a few examples).

During the last 15–20 years, there have been dramatic – and surprising – developments in SAT solving technology that have improved real-world performance by many orders of magnitude. But perhaps even more surprisingly, the best SAT solvers today are still based on relatively simple methods from the early 1960s, searching for proofs in the so-called resolution proof system. While such solvers can often handle formulas with millions of variables, there are also known tiny formulas with just a few hundred variables that cause even the very best solvers to stumble. The fundamental question of when SAT solvers perform well or badly, and what underlying properties of the formulas influence SAT solver performance, remains very poorly understood. Other practical SAT solving issues, such as how to optimize memory management and how to exploit parallelization on modern multicore architectures, are even less well studied and understood from a theoretical point of view.

Another intriguing fact is that although other mathematical methods of reasoning are known that are much stronger than resolution in theory, in particular methods based on algebra and geometry, attempts to harness the power of such methods have failed to deliver any significant improvements in practical performance – indeed, such solvers often struggle even to match the performance of resolution-based solvers. And while resolution is a fairly well-understood proof system, even very basic questions about these stronger algebraic and geometric methods remain wide open.

We believe that computational complexity can shed light on the power and limitations on current and possible future SAT solving techniques, and that problems encountered in SAT solving can spawn interesting new areas in theoretical research. We see great potential for interdisciplinary research at the border between theory and practice in this area, and believe that more vigorous interaction between practitioners and theoreticians could have major long-term impact in both academia and industry.

Goals of the Workshop

A strong case can be made for the importance of increased exchange between the two fields of SAT solving on the one hand and proof complexity (and more broadly computational complexity) on the other. While the two areas have enjoyed some exchanges, it seems fair to say that there has been relatively low level of interaction, given how many questions would seem to be of mutual interest. Below, we try to outline some such questions that served as motivation for organizing this seminar. We want to stress that this list is far from exhaustive, and in fact we believe one important outcome of the seminar was to stimulate the process of uncovering other questions of common interest.

What Makes Formulas Hard or Easy in Practice for Modern SAT Solvers?

The best SAT solvers known today are based on the DPLL procedure, augmented with optimizations such as conflict-driven clause learning (CDCL) and restart strategies. The propositional proof system underlying such algorithms, resolution, is arguably the most well-studied system in all of proof complexity.

Given the progress during the last decade on solving large-scale instances, it is natural to ask what lies behind the spectacular success of CDCL solvers at solving these instances. And given that there are still very small formulas that resist even the most powerful CDCL solvers, a complementary interesting question is if one can determine whether a particular formula is hard or tractable. Somewhat unexpectedly, very little turns out to be known about these questions.

In view of the fundamental nature of the SAT problem, and in view of the wide applicability of modern SAT solvers, this seems like a clear example of a question of great practical importance where the theoretical field of proof complexity could potentially provide useful insights. In particular, one can ask whether one could find theoretical complexity measures for formulas that would capture the practical hardness of these formulas in some nice and clean way. Besides greatly advancing our theoretical understanding, answering such a question could also have applied impact in the longer term by clarifying the limitations, and potential for further improvements, of modern SAT solvers.

Can Proof Complexity Shed Light on Crucial SAT Solving Issues?

Understanding the hardness of proving formulas in practice is not the only problem for which more applied researchers would welcome contributions from theoretical computer scientists. Examples of some other possible practical questions that would merit from a deeper theoretical understanding follow below.

- Firstly, we would like to study the question of memory management. One major concern for clause learning algorithms is to determine how many clauses to keep in memory. Also, once the algorithm runs out of the memory currently available, one needs to determine

which clauses to throw away. These questions can have huge implications for performance, but are poorly understood.

- In addition to clause learning, the concept of restarts is known to have decisive impact on the performance on modern CDCL solvers. It would be nice to understand theoretically why this is so. The reason why clause learning increases efficiency greatly is clear – without it the solver will only generate so-called tree-like proofs, and tree-like resolution is known to be exponentially weaker than general resolution. However, there is still ample room for improvement of our understanding of the role of restarts and what are good restart strategies.
- Given that modern computers are multi-core architectures, a highly topical question is whether this (rather coarse-grained) parallelization can be used to speed up SAT solving. Our impression is that this is an area where much practical work is being carried out, but where comparatively little theoretical study has been done. Thus, the first step here would consist of understanding what are the right questions to ask and coming up with a good theoretical framework for investigating them.

While there are some successful attempts in parallelizing SAT, obtained speed-ups are rather modest. This is a barrier for further adoption of SAT technology already today and will become a more substantial problem as thousands of cores and cloud computing are becoming the dominant computing platforms. A theoretical understanding on how SAT can be parallelized will be essential to develop new parallelization strategies to adapt SAT to this new computing paradigm.

Can we build SAT Solvers based on Stronger Proof Systems than Resolution?

Although the performance of modern CDCL SAT solvers is impressive, it is nevertheless astonishing, not to say disappointing, that the state-of-the-art solvers are still based on simple resolution. Resolution lies very close to the bottom in the hierarchy of propositional proof systems, and there are many other proof systems based on different forms of mathematical reasoning that are known to be strictly stronger. Some of these appear to be natural candidates for serving as a basis for stronger SAT solvers than those using CDCL.

In particular, proof systems such as polynomial calculus (based on algebraic reasoning) and cutting planes (based on geometry) are known to be exponentially more powerful than resolution. While there has been some work on building SAT solvers on top of these proof systems, progress has been fairly limited. As part of the seminar, we invited experts on algebraic and geometric techniques to discuss what the barriers are that stop us from building stronger algebraic or geometric SAT solvers, and what is the potential for future improvements. An important part of this work would seem to be to gain a deeper theoretical understanding of the power and limitations of these proof methods. Here there are a number of fairly long-standing open theoretical questions. At the same time, only in the last couple of years proof complexity has made substantial progress, giving hope that the time is ripe for decisive break-throughs in these areas.

Organization of the Workshop

The scientific program of the seminar consisted of 26 talks. Among these there were five 80-minute tutorials on core topics of the seminar:

- proof complexity (Paul Beame),
- conflict-driven clause learning (CDCL) SAT solvers (João Marques-Silva),

- proof systems connected to SAT solving (Sam Buss),
- preprocessing and inprocessing (Matti Järvisalo),
- SAT and SMT (Nikolaj Bjørner).

Throughout, the tutorials were well-received as a means of introducing the topics and creating a common frame of reference for participants from the different communities.

There were also nine slightly shorter survey talks of 50 minutes which were intended to give overviews of a number of important topics for the seminar:

- semialgebraic proof systems (Albert Atserias),
- pseudo-Boolean constraints and CDCL (Daniel Le Berre),
- Gröbner bases (Manuel Kauers),
- SAT-enabled verification of state transition systems, (Karem Sakallah),
- SAT and computational complexity (Ryan Williams)
- the (strong) exponential time hypothesis and consequences (Ryan Williams),
- SAT and parameterized complexity (Stefan Szeider),
- QBF solving (Nina Narodytska),
- random satisfiability (Dimitris Achlioptas).

Most tutorials and survey talks were scheduled early in the week, to create a conducive atmosphere for collaboration on open problems later in the week. The rest of the talks were 25-minute presentations on recent research of the participants. The time between lunch and afternoon coffee was left for self-organized collaborations and discussions, and there was no schedule on Wednesday afternoon.

Based on polling of participants before the seminar week, it was decided to have an open problem session on Monday evening, and on Wednesday evening there was a panel discussion. The organizing committee also considered the option of having a poster session to give more researchers the opportunity to present recent research results, but the feedback in the participant poll was negative and so this idea was dropped.

2 Table of Contents

Executive Summary

Armin Biere, Vijay Ganesh, Martin Grohe, Jakob Nordström, and Ryan Williams 99

Overview of Presentations

An introduction to proof complexity <i>Paul Beame</i>	106
Tutorial on conflict-driven clause learning (CDCL) SAT solvers <i>João Marques-Silva</i>	106
An Introduction to Semialgebraic Proofs: Basic Definitions and Results <i>Albert Atserias</i>	106
Handling Pseudo-Boolean constraints in a CDCL solver: a practical survey <i>Daniel Le Berre</i>	107
Gröbner bases <i>Manuel Kauers</i>	107
Tutorial on proof systems connected to SAT solving <i>Sam Buss</i>	107
Tutorial on preprocessing and inprocessing <i>Matti Järvisalo</i>	107
An Empirical Understanding of Conflict-Driven Clause-Learning SAT Solvers <i>Vijay Ganesh</i>	108
MaxSAT Solving with SAT Oracles <i>João Marques-Silva</i>	108
SAT-Enabled Verification of State Transition Systems <i>Karem Sakallah</i>	109
Machine learning for SAT <i>Holger Hoos</i>	109
How SAT Solvers Could (And Do) Prove Lower Bounds + (S)ETH and A survey of Consequences <i>Ryan Williams</i>	109
A Survey on Parameterized Complexity and SAT <i>Stefan Szeider</i>	110
From SAT to SMT – a Tutorial <i>Nikolaj S. Bjørner</i>	110
Survey on QBF solving <i>Nina Narodytska</i>	110
QBF proof complexity <i>Olaf Beyersdorff</i>	111
Parallel SAT Solving or To Share or Not To Share <i>Armin Biere</i>	111
Linear Temporal Logic Satisfiability Checking <i>Kristin Yvonne Rozier</i>	111

Resolution Proofs of Bounded Width <i>Christoph Berkholz</i>	112
An Ultimate Trade-Off in Propositional Proof Complexity <i>Alexander Razborov</i>	112
Narrow Proofs May Be Maximally Long <i>Massimo Lauria</i>	113
A Survey of Random Satisfiability <i>Dimitris Achlioptas</i>	113
Space and Random CNFs <i>Ilario Bonacina</i>	113
Improving and Evaluating a Hybrid Approach to Max-SAT Solving <i>Jessica Davies</i>	113
Bit-Vectors: Complexity and Decision Procedures <i>Andreas Fröhlich</i>	114
Some Open Problems	
Minimum variable space and minimum depth of resolution refutations <i>Alexander Razborov</i>	114
Exact counting for k-SAT <i>Ryan Williams</i>	115
Optimality of Regular Resolution? <i>Alasdair Urquhart</i>	115
How and why does VSIDS work? (Full simulation of resolution by CDCL with heuristics?) <i>Alexandra Goultiaeva, Armin Biere, and Vijay Ganesh</i>	115
Learning definitions through extended resolution <i>Armin Biere</i>	116
Limits of portfolio based parallel SAT solving <i>Armin Biere</i>	116
What is the relationship, if any, between cluster analysis and survey propagation on application SAT instances? <i>Allen Van Gelder</i>	116
Why does conflict-driven search work so well? (How do CDCL solvers exploit the structure of real-world instances?) <i>Karem Sakallah, Sharad Malik, and Vijay Ganesh</i>	117
How to cut directed paths in a dag (related to the complexity of CircuitSAT) <i>Edward A. Hirsch</i>	117
How Total Space and Monomial Space relate with other complexity measures? <i>Ilario Bonacina</i>	117
Random <i>k</i> -SAT <i>Dimitris Achlioptas</i>	118
The complexity of the parity principle in semi-algebraic systems <i>Paul Beame</i>	118

Panel Discussion 119

Examples of Outcomes of the Workshop 119

Evaluation by Participants 120

Participants 122

3 Overview of Presentations

In this section we list the titles and abstracts of all presentations given during the seminar.

3.1 An introduction to proof complexity

Paul Beame (University of Washington – Seattle, US)

License  Creative Commons BY 3.0 Unported license
© Paul Beame

We give an overview of proof complexity including its basic definitions and many examples of natural and widely-studied propositional proof systems including inference systems using logical formulas, circuits, polynomials, and linear and polynomial inequalities. We also show how every complete SAT solver also yields a propositional proof system. We describe many of the known relationships between propositional proof systems and known bounds on their efficiency. We show some of the key techniques for bounding the lengths of propositional proofs, including relationships between their size, width, and degree and we show how this is related to forms of graph expansion of their input formulas. Finally, we describe a number of classes of natural examples of formulas that are hard to prove in various proof systems.

3.2 Tutorial on conflict-driven clause learning (CDCL) SAT solvers

João Marques-Silva (INESC-ID – Lisboa, PT)

License  Creative Commons BY 3.0 Unported license
© João Marques-Silva

Conflict-driven clause learning (CDCL) SAT solvers represent the de facto standard solver in practical problem solving with SAT, being used in the most visible and most successful practical applications of SAT. This tutorial will give an overview of the key concepts and techniques used in modern CDCL SAT solvers.

3.3 An Introduction to Semialgebraic Proofs: Basic Definitions and Results

Albert Atserias (UPC – Barcelona, ES)

License  Creative Commons BY 3.0 Unported license
© Albert Atserias

Boolean satisfiability is a special case of integer linear programming, so we can hope to integrate some of their methods to SAT solvers. We will go over well-studied semialgebraic techniques, namely Gomory-Chvátal cuts and lift-and-project methods, and present some cases where they beat a resolution-based approach, as well as some lower bounds.

3.4 Handling Pseudo-Boolean constraints in a CDCL solver: a practical survey

Daniel Le Berre (CNRS – Lens, FR)

License  Creative Commons BY 3.0 Unported license
© Daniel Le Berre

CDCL solvers have been quickly extended to handle arbitrary constraints. Doing so while preserving the original proof system of the solver does not require much changes to the solver.

Extending the proof system of the solver is however much more challenging. The talk will emphasize the extension of the CDCL architecture to the so called “generalized resolution” proof system, which lies between resolution and cutting planes proof systems, to handle Pseudo-Boolean constraints.

It will especially point out the strong requirements of the CDCL architecture on the proof system used for conflict analysis. Gory details about the constraints derived by such extended CDCL solver on benchmarks such as pigeon hole formulas will highlight both the strength and weaknesses of the resulting solver.

3.5 Gröbner bases

Manuel Kauers (Universität Linz, AT)

License  Creative Commons BY 3.0 Unported license
© Manuel Kauers

We explain what Gröbner bases are, why they are interesting, and how they are computed. The focus of the talk is on computational aspects. We will therefore not say much about how Gröbner bases can be used for solving all sorts of problems in commutative algebra. Instead, after discussing the classical Buchberger algorithm for computing Gröbner basis, we will try to sketch the underlying ideas of more recent algorithms.

3.6 Tutorial on proof systems connected to SAT solving

Sam Buss (University of California – San Diego, US)

License  Creative Commons BY 3.0 Unported license
© Sam Buss

Most SAT solvers implicitly generate refutation in the resolution proof system. We review this connection and characterize the shape of proofs generated by a CDCL solver. We introduce proof systems weaker than resolution that model these proofs.

3.7 Tutorial on preprocessing and inprocessing

Matti Järvisalo (University of Helsinki, FI)

License  Creative Commons BY 3.0 Unported license
© Matti Järvisalo

This tutorial aims at covering (i) some of the most important preprocessing techniques used today in practice in conjunction with SAT solvers, and (ii) a generic “inprocessing” proof

system capturing the deductions made by inprocessing SAT solvers that interleave CDCL search and preprocessing steps during search.

3.8 An Empirical Understanding of Conflict-Driven Clause-Learning SAT Solvers

Vijay Ganesh (University of Waterloo, CA)

License  Creative Commons BY 3.0 Unported license
© Vijay Ganesh

Modern conflict-driven clause-learning (CDCL) Boolean SAT solvers routinely solve very large industrial SAT instances in relatively short periods of time. This phenomenon has stumped both theoreticians and practitioners since Boolean satisfiability is an NP-complete problem widely believed to be intractable. It is clear that these solvers somehow exploit the structure of real-world instances. However, to-date there have been few results that precisely characterize this structure, or shed any light on why these SAT solvers are so efficient.

In this talk, I will present results that provide a deeper empirical understanding of why CDCL SAT solvers are so efficient. First, we provide evidence that industrial SAT instances have “good community structure”, and that this correlates more strongly with the running time of SAT solvers than traditional complexity-theoretic measures of SAT instance size such as number of clauses, variables or clause-variable ratio. Second, we characterize the famous VSIDS branching heuristic through a set of behavioral invariants that we discovered through a rigorous scientific process. These invariants include the following: First, VSIDS picks high-centrality bridge variables in the community structure of SAT instances much more often than other variables. Second, the multiplicative decay in VSIDS acts as an exponential moving average (EMA). Third, VSIDS is spatially and temporal focused (localized) with respect to the community structure of the SAT instance. We believe that the net effect of these behaviors of VSIDS is that it essentially enables the CDCL SAT solver to carry out a divide-and-conquer strategy by separating and then solving the communities of an instance.

Finally, I will present an abstract model of a SAT solver as an “active learner with deductive corrective feedback” that we believe is an accurate and analyzable mathematical model of CDCL solvers. I will also provide evidence that many successful techniques in formal verification and, more broadly, in software engineering can be abstractly modeled as “reinforcement learners with deductive corrective feedback”.

3.9 MaxSAT Solving with SAT Oracles

João Marques-Silva (INESC-ID – Lisboa, PT)

License  Creative Commons BY 3.0 Unported license
© João Marques-Silva

Given an unsatisfiable formula, the maximum satisfiability problem (MaxSAT) is to identify a maximal subset of clauses that can be simultaneously satisfied. MaxSAT finds a growing number of practical applications, that include fault localization in software, design debugging in hardware, different applications in bioinformatics, timetabling and scheduling problems, among many others. For practical purposes, the most effective algorithms are based on iterative identification and relaxation of unsatisfiable subformulas using SAT solvers as oracles. This talk gives a brief overview of MaxSAT algorithms based on SAT oracles, and highlights what are currently the most effective techniques.

3.10 SAT-Enabled Verification of State Transition Systems

Karem A. Sakallah (University of Michigan – Ann Arbor, US)

License © Creative Commons BY 3.0 Unported license
© Karem Sakallah

The sequential behavior of complex artifacts, such as a hardware design or a software programs, is commonly captured by modeling the artifact as a formal state transition system. Given a desired (safety) property on the states of such a system, an important verification challenge is to determine whether all states reachable from a given (safe) initial state are safe and, if not, to produce an execution trace leading from the initial state to an unsafe state. Algorithmic approaches for solving this problem, in contrast to interactive theorem proving or proof checking methods, are what is referred to in the literature as model checking (MC).

In this talk I will briefly survey the evolution of MC over the last 30+ years highlighting the critical role SAT technology played in scaling MC to transition systems with exponentially-sized state spaces. I will also describe two specific applications, one in hardware and one in software, to illustrate the architecture of a scalable SAT-based verification environment.

3.11 Machine learning for SAT

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

License © Creative Commons BY 3.0 Unported license
© Holger Hoos

In this presentation I will explain how machine learning methods can be used to automatically configure, select, combine and assess SAT solvers. I will briefly cover algorithm configuration techniques, such as SMAC (as used in the recent Configurable SAT Solver Challenges), automated algorithm selectors, such as SATzilla, automatic techniques for constructing parallel solver portfolios and finally, an interesting approach for assessing the scaling of solver performance with instance size that recently produced evidence that SLS-based SAT solvers like WalkSAT have running time polynomial in instance size for phase transition random-3-SAT instances.

3.12 How SAT Solvers Could (And Do) Prove Lower Bounds + (S)ETH and A survey of Consequences

Ryan Williams (Stanford University, US)

License © Creative Commons BY 3.0 Unported license
© Ryan Williams

This is a merger of two tutorial talks: one by me on SAT algorithms and connections to computational complexity theory, and one by Mohan (cancelled) on the Exponential Time Hypothesis and the Strong Exponential Time Hypothesis.

3.13 A Survey on Parameterized Complexity and SAT

Stefan Szeider (TU Wien, AT)

License  Creative Commons BY 3.0 Unported license
© Stefan Szeider

In this talk I will discuss basic concepts of parameterized complexity (such as fixed-parameter tractability, reductions, hardness, and kernelization) and survey parameterized complexity results related to satisfiability (SAT). The focus will be on laying out what kind of questions can be asked and not on technical details.

3.14 From SAT to SMT – a Tutorial

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

License  Creative Commons BY 3.0 Unported license
© Nikolaj S. Bjørner

Satisfiability Modulo Theories (SMT) solvers are used in many modern program verification, analysis and testing tools. They owe their scale and efficiency thanks to advances in search algorithms underlying modern SAT solvers and first-order theorem provers. They owe their versatility in software development applications thanks to specialized algorithms supporting theories, such as numbers and algebraic data-types, of relevance for software engineering.

This tutorial introduces algorithmic principles of SMT solving, taking as basis modern SAT solvers and integration with specialized theory solvers and quantifier reasoning. We detail some of the algorithms used for main theories used in current SMT solvers and survey newer theories and approaches to integrating solvers. The tutorial also outlines some application scenarios where SMT solvers have found use, including program verification, network analysis, symbolic model checking, test-case generation, and white-box fuzzing.

3.15 Survey on QBF solving

Nina Narodytska (Carnegie Mellon University, US)

License  Creative Commons BY 3.0 Unported license
© Nina Narodytska

Quantified Boolean formulas are a natural extension of propositional formulas with universal and existential quantifiers. QBF solvers are used in solving many problems in knowledge representation and reasoning, automated planning, and computer aided design.

In this talk, I will introduce the QBF problem and survey state-of-the-art techniques used in QBF solving. Then I will focus on a recent and successful approach that is based on the counterexample-guided abstraction refinement (CEGAR) paradigm. This approach proved very effective on a large number of industrial families of benchmarks.

3.16 QBF proof complexity

Olaf Beyersdorff (University of Leeds, GB)

License  Creative Commons BY 3.0 Unported license
© Olaf Beyersdorff

In this talk we give an overview of the relatively young field of QBF proof complexity. We explain the main resolution-based proof systems for QBF, modelling CDCL and expansion-based solving. As our main contribution we exhibit a new and elegant proof technique for showing lower bounds in QBF proof systems based on strategy extraction. This technique provides a direct transfer of circuit lower bounds to lengths of proofs lower bounds. We use our method to show the hardness of a natural class of parity formulas for Q-resolution. Our lower bounds imply new exponential separations between two different types of resolution-based QBF calculi: proof systems for CDCL-based solvers and proof systems for expansion-based solvers. The relations between proof systems from the two different classes were not known before.

3.17 Parallel SAT Solving or To Share or Not To Share

Armin Biere (Universität Linz, AT)

License  Creative Commons BY 3.0 Unported license
© Armin Biere

We give a brief introduction into the problem and the current state-of-the-art of parallel SAT solving, mostly from a practical point of view. The talk continues with discussing current challenges.

3.18 Linear Temporal Logic Satisfiability Checking

Kristin Yvonne Rozier (University of Cincinnati, US)

License  Creative Commons BY 3.0 Unported license
© Kristin Yvonne Rozier

Formal verification techniques are growing increasingly vital for the development of safety-critical software and hardware. Techniques such as requirements-based design and model checking have been successfully used to verify systems for air traffic control, airplane separation assurance, autopilots, logic designs, medical devices, and other functions that ensure human safety. Formal behavioral specifications written early in the system-design process and communicated across all design phases increase the efficiency, consistency, and quality of the system under development. We argue that to prevent introducing design or verification errors, it is crucial to test specifications for satisfiability.

In 2007, we established LTL satisfiability checking as a sanity check: each system requirement, its negation, and the set of all requirements should be checked for satisfiability before being utilized for other tasks, such as property-based system design or system verification via model checking. We demonstrated that LTL satisfiability checking reduces to model checking; an extensive experimental evaluation proved that for LTL satisfiability checking, the symbolic approach is superior to the explicit approach. However, the performance of the symbolic

approach critically depends on the encoding of the formula. Since 1994, there had been essentially no new progress in encoding LTL formulas as symbolic automata for BDD-based analysis. We introduced a set of 30 symbolic automata encodings, demonstrating that a portfolio approach utilizing these encodings translates to significant, sometimes exponential, improvement over the standard encoding for symbolic LTL satisfiability checking. In recent years, LTL satisfiability checking has taken off, with others inventing exciting new methods to scale with increasingly complex systems. We revisit the benchmarks for LTL satisfiability checking that have become the de facto industry standard and examine the encoding methods that have led to leaps in performance. We highlight the past and present, and look to the future of LTL satisfiability checking, a sanity check that now has an established place in the development cycles of safety-critical systems.

3.19 Resolution Proofs of Bounded Width

Christoph Berkholz (KTH Royal Institute of Technology, SE)

License  Creative Commons BY 3.0 Unported license
© Christoph Berkholz

The talk focuses on the structure and complexity of resolution refutations of bounded width (where every clause contains at most k literals).

Such refutations can be found in time $n^{O(k)}$ by exhaustively deriving all possible clauses with at most k literals. We show that this upper bound is tight by proving a matching lower bound. Furthermore, deciding whether there exists a resolution refutation of bounded width is EXPTIME-complete, whereas the same problem for regular resolution is PSPACE-complete.

We will also discuss the structure of bounded width refutations in terms of classical proof complexity measures such as resolution depth, (treelike) resolution size and clause space.

3.20 An Ultimate Trade-Off in Propositional Proof Complexity

Alexander Razborov (University of Chicago, US)

License  Creative Commons BY 3.0 Unported license
© Alexander Razborov

Trade-off results in complexity theory follow this general pattern: a task is exhibited that is easy with respect to a chosen complexity measure but becomes much harder after requiring that the protocol is efficient with respect to another, normally very different, measure. In most cases, “much harder” means “as hard as an average task of comparable size” without imposing any restrictions on the protocol.

In this talk we exhibit an unusually strong trade-off result between width and tree-like resolution proof size that significantly deviates from this pattern. Namely, we construct unsatisfiable k -CNFs that possess refutations of very small width $O(k)$ but such that any tree-like resolution refutation of even mildly sublinear width $n^{1-\epsilon}/k$ must be of double exponential size $\exp(n^{\Omega(k)})$. This is exponentially larger than the trivial 2^n size bound to which all unsatisfiable CNFs with n variables are entitled.

3.21 Narrow Proofs May Be Maximally Long

Massimo Lauria (KTH Royal Institute of Technology, SE)

License  Creative Commons BY 3.0 Unported license
© Massimo Lauria

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

3.22 A Survey of Random Satisfiability

Dimitris Achlioptas (University of California – Santa Cruz, US)

License  Creative Commons BY 3.0 Unported license
© Dimitris Achlioptas

Given a CNF formula F , let $S(F)$ denote its set of satisfying assignments. We consider a random k -CNF formula F on n variables, constructed by adding m random clauses one by one, each clause selected uniformly at random among all $2^k \binom{n}{k}$ possible clauses. The talk will give a survey of results about random satisfiability by narrating the “video” of $S(F)$ as clauses are added. We will see that two important phase transitions occur (neither of which is the satisfiability transition) and emphasis will be placed on their potential algorithmic implications. No familiarity with random satisfiability will be assumed.

3.23 Space and Random CNFs

Ilario Bonacina (University of Rome “La Sapienza”, IT)

License  Creative Commons BY 3.0 Unported license
© Ilario Bonacina

We will see some space lower bounds in Resolution and Polynomial Calculus Resolution (PCR) for random k -CNFs. More precisely about random 3-CNFs: a quadratic lower bound for the total space needed in Resolution to refute such formulas and a linear lower bound for monomial space in PCR.

3.24 Improving and Evaluating a Hybrid Approach to Max-SAT Solving

Jessica Davies (IST Austria – Klosterneuburg, AT)

License  Creative Commons BY 3.0 Unported license
© Jessica Davies

MaxHS is a recent approach to solving Max-SAT that utilizes a hybrid algorithm that exploits both a SAT solver and an IP solver as black-boxes. This approach has a number of attractive

properties, but in the recent Max-SAT Evaluations it has not performed as well as other purely SAT-based solvers. In this paper we examine a current implementation of MaxHS and find a number of improvements. With these improvements implemented we compare the performance of the approach to other approaches for solving Max-SAT. Our results indicate that the hybrid approach remains a promising direction for further research.

3.25 Bit-Vectors: Complexity and Decision Procedures

Andreas Fröhlich (Universität Linz, AT)

License  Creative Commons BY 3.0 Unported license
© Andreas Fröhlich

Bit-vectors are important for many practical applications in verification. We discuss theory and practice by giving complexity results and presenting several alternative decision procedures.

4 Some Open Problems

Before the seminar, the organizers collected a list of open problems from the participants that could potentially be discussed during the open problem session Monday evening and at other times during the week. All submitted problems were collected at the webpage <http://www.csc.kth.se/~jakobn/dagstuhl15171/openproblems.php>. Many of these problems were indeed discussed during the Monday evening problem session, and in addition other problems were raised there as well.

Below follows a hopefully representative selection of these open problems. The list is basically unsorted except it is (roughly) in chronological order of submission. Some partially overlapping problems have been merged. The full list of problems is still available at <http://www.csc.kth.se/~jakobn/dagstuhl15171/openproblems.php>. One suggestion put forward during the seminar week was to collect these and other research problems on a wiki-style website to stimulate research. This seems like a very attractive idea, and is something that might be implemented in the future.

4.1 Minimum variable space and minimum depth of resolution refutations

Alexander Razborov (University of Chicago, US)

License  Creative Commons BY 3.0 Unported license
© Alexander Razborov

Can it be the case that minimum variable space is equivalent, up to a polynomial and $\log n$ factors, to the minimum depth of resolution refutations? This is true if we additionally normalize variable space by log of the proof length, therefore an equivalent form of our question is this: does there exist a strong ultimate tradeoff between variable space and proof length?

4.2 Exact counting for k-SAT

Ryan Williams (Stanford University, US); (originally from Rahul Santhanam)

License © Creative Commons BY 3.0 Unported license
© Ryan Williams

It is known that k -SAT with n variables and m clauses can be solved in about $2^{n-n/O(k)} \cdot \text{poly}(m)$ time, and these are the best known running times for the worst case. For computing the number of k -SAT solutions, there is a randomized algorithm of Impagliazzo, Matthews, and Paturi (SODA'12) running in $2^{n-n/O(k)} \cdot \text{poly}(m)$ time, and a deterministic algorithm of Beame, Impagliazzo, and Srinivasan (CCC'12) running in worse time.

Is there a deterministic worst-case $\#k$ -SAT algorithm running in $2^{n-n/O(k)} \cdot \text{poly}(m)$ time? (Give an algorithm, or evidence against its existence.)

4.3 Optimality of Regular Resolution?

Alasdair Urquhart (University of Toronto, CA)

License © Creative Commons BY 3.0 Unported license
© Alasdair Urquhart

- Show that for well known examples such as the pigeonhole principle (PHP) and Tseitin formulas, regular resolution is optimal. This conjecture seems very plausible to me, but I don't see how to approach it at the moment.
- More generally, you can ask: Can you give general conditions on a set of clauses that ensure that regular resolution is optimal? In general, the examples separating general and unrestricted resolution have a rather artificial appearance, where we add "spoiler variables" to mess up any regular refutation.
- A closely related problem that may be more accessible is this: for the same set of examples, show that the regular width and the unrestricted width of a refutation are the same. Are there general conditions that ensure this equality?

4.4 How and why does VSIDS work? (Full simulation of resolution by CDCL with heuristics?)

Alexandra Goultiaeva (Google Waterloo, CA), Armin Biere (Universität Linz, AT), and Vijay Ganesh (University of Waterloo, CA)

License © Creative Commons BY 3.0 Unported license
© Alexandra Goultiaeva, Armin Biere, and Vijay Ganesh

The variable scoring scheme VSIDS (variable state independent decaying sum) introduced by Chaff and its modern variants is crucial for the speed of CDCL solvers. There is almost no empirical investigation on how it really works, and further no theoretical explanation why it is working.

In particular, it has been proven that CDCL SAT solvers p-simulate resolution. The order of decisions is assumed to be arbitrary, i.e., the proof shows that (if a short resolution proof exists) there exists a sequence of decisions that would allow the solver to find a short resolution proof. I.e, a result that either:

- shows that whenever a short resolution proof exists, there will always be a sequence of decisions that respects VSIDS ordering and allows the solver to find a short resolution proof, or
- shows a counterexample where a short resolution proof exists but a solver respecting VSIDS ordering (regardless of tie-breaking) can never find a short proof.

4.5 Learning definitions through extended resolution

Armin Biere (Universität Linz, AT)

License  Creative Commons BY 3.0 Unported license
© Armin Biere

There has been attempts to shrink learned clauses by introducing definitions in the sense of extended resolutions, which however in practice has not really been effective. It is unclear whether these newly introduced literals can really be used in the search process and shrink proofs. The question is whether it is possible to come up with a more general but practical scheme to introduce definitions, which allow to shrink proof size and improve SAT solving in practice too.

4.6 Limits of portfolio based parallel SAT solving

Armin Biere (Universität Linz, AT)

License  Creative Commons BY 3.0 Unported license
© Armin Biere

Portfolio based SAT solving is the dominating approach in the parallel application track of the SAT competition. However, the improvements we saw in the last two years are apparently based on using better sharing schemes for learned clauses, thus kind of implicit work splitting. From a practical point of view it is first of all still unclear how much of the success of solvers like Penelope or Plingeling can be contributed to the portfolio idea and how much is due to splitting the work. As the number of compute units is increased it is conjectured that the relative contribution of the portfolio part will saturate. Does this happen and when?

4.7 What is the relationship, if any, between cluster analysis and survey propagation on application SAT instances?

Allen Van Gelder (University of California – Santa Cruz, US)

License  Creative Commons BY 3.0 Unported license
© Allen Van Gelder

There are several recent works on cluster analysis (AKA community structure) of application SAT instances. They seem to focus on connections between clause learning, VSIDS, and the page-rank algorithm.

- What new idea is needed for survey propagation to be useful on application SAT instances?
- Is survey propagation somehow useful on unsatisfiable application SAT instances? Can certain behavior suggest the application SAT instance is unsat and give evidence?
- Can cluster analysis on application SAT instances give a hint or prediction whether the application SAT instance is unsat or sat?

4.8 Why does conflict-driven search work so well? (How do CDCL solvers exploit the structure of real-world instances?)

Karem Sakallah (University of Michigan – Ann Arbor, US), Sharad Malik (Princeton University, US), and Vijay Ganesh (University of Waterloo, CA)

License  Creative Commons BY 3.0 Unported license
© Karem Sakallah, Sharad Malik, and Vijay Ganesh

Empirical evaluation of solver performance suggests that the two most important features of modern SAT solvers are *conflict-driven clause learning* and *conflict-driven branching*. Tracing the execution of a modern conflict-driven solver seems to show that the solver is aggressively trying to falsify the formula (looking for conflicts) and only when that fails does it yield a satisfying assignment. This strategy seems to work quite well on a very diverse set of benchmarks. Why? Can we characterize when it does not work? What problem structure causes an aggressive falsification approach to fail? What other strategies can we envision to complement conflict-driven search?

4.9 How to cut directed paths in a dag (related to the complexity of CircuitSAT)

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

License  Creative Commons BY 3.0 Unported license
© Edward A. Hirsch

Consider directed acyclic graphs with vertices of indegree at most two (that is, Boolean circuits). Prove (or disprove) that for every $\epsilon > 0$ there is a constant $K = K(\epsilon)$ such that for every n large enough in every such dag with n vertices there is a subset of vertices of size at most $\epsilon \cdot n$ such that its removal (with incident edges) leaves no directed paths of length more than K .

4.10 How Total Space and Monomial Space relate with other complexity measures?

Ilario Bonacina (University of Rome “La Sapienza”, IT)

License  Creative Commons BY 3.0 Unported license
© Ilario Bonacina

Given an unsatisfiable CNF ϕ let's see a refutation of it in Resolution (res. PCR) as a sequence of memory configurations, i.e. set of clauses (res. polynomials) such that each memory configuration is obtained from the previous one either (i) removing some clause (resp. polynomial), or (ii) adding some clause from ϕ , or (iii) inferring some consequence applying the inference rules to something in memory.

$Mspace_{PCR}(\phi \vdash \perp) \geq m$ means that for every PCR refutation π of ϕ (according to the previous model) there must be some memory configuration in π in which at least m distinct monomials appear (maybe in several places in the polynomials in that memory configuration).

$TSpace(\phi \vdash \perp) \geq m$ means that for every (Res/PCR) refutation π of ϕ (according to the previous model) there must be some memory configuration in π in which the total number of occurrences of literals in that memory configuration is at least m .

So the questions are the following:

- Is it the case that given a k -CNF ϕ , $TSpace_{RES}(\phi \vdash \perp) = \Omega((width(\phi \vdash \perp) - k)^2)$?
- Is it the case that given a k -CNF ϕ , $MSpace_{PCR}(\phi \vdash \perp) = \Omega(degree(\phi \vdash \perp) - k)$?
- Is there any k -CNF ϕ in n variables and $n^{O(1)}$ clauses such that $TSpace_{PCR}(\phi \vdash \perp) = \Omega(n^2)$? It should be true w.h.p. for random k -CNFs for any $k \geq 3$ and with clause density a constant above the unsatisfiability threshold.

4.11 Random k -SAT

Dimitris Achlioptas (University of California – Santa Cruz, US)

License  Creative Commons BY 3.0 Unported license
© Dimitris Achlioptas

- The sat/unsat threshold for random 10-SAT is provably > 700 . Solve random 10-SAT instances with 100,000 variables of density 600 (or greater). (hard)
- The sat/unsat threshold for random 6-SAT is provably > 40 . Solve random 6-SAT instances with 100,000 variables of density 35 (or greater). (not easy)
- The mixture of $(1 - \epsilon)n$ random 2-clauses and $(2/3)n$ random 3-clauses (on the same variables) is satisfiable with high probability, for every $\epsilon > 0$. Prove that $2/3$ is best possible. That is, prove that for every $\delta > 0$, there exists $\epsilon > 0$ such that such a mixture is unsatisfiable. (hard)

4.12 The complexity of the parity principle in semi-algebraic systems

Paul Beame (University of Washington – Seattle, US)

License  Creative Commons BY 3.0 Unported license
© Paul Beame

Determine the complexity of the parity principle (also known as the mod 2 counting principle, or the matching principle on K_{2n+1}) in semi-algebraic systems, especially LS and LS+:

This has a variable for each edge of the complete graph on an odd number of vertices. In clausal form this has clauses like the bijective pigeonhole for each vertex but it is easy to derive $\sum_{i \neq j} x_{ij} = 1$ in small size in these systems. (In LS it takes degree $\Omega(n)$ to derive this but it is only quadratic size. In LS+ there is a rank one derivation.)

In cutting planes it is easy to derive a contradiction from this since one can add all of the equations to get $2 \sum_{i,j:i \neq j} x_{ij} = 2n + 1$ and rounding in both directions yields a contradiction. However, it is not clear how to simulate this “division by 2” in any semi-algebraic system. This is related to the Knapsack problem considered by Grigoriev. He showed that if a sum of m variables is an odd number that is near the middle of the interval $[0, m]$ then Positivstellensatz degree is large. Using the methods of Kojevnikov and Itsyksen this yields tree-like size lower bounds for LS. The differences here are that there are $\binom{2n+1}{2}$ variables and the $2n + 1$ bound is nowhere near the middle of the range $[0, \binom{m}{2}]$, and we have separate equations for subset of the variables.

5 Panel Discussion

On Wednesday evening there was a panel discussion with Paul Beame, Nikolaj Bjørner, Sam Buss, Sharad Malik, Karem Sakallah, and Stefan Szeider serving as members of the panel. The panel members opened the discussion with a short “keynote remark” each (of around 4–5 minutes), after which followed a discussion of a little bit more than an hour between panel members and all participants present. The purpose of the panel was to discuss question such as promising and/or important future research directions, how (and if) we should get more interaction between practitioners and theoreticians doing SAT-related research, or whatever else the seminar participants wanted to talk about.

The panel discussion brought out several socio-scientific issues at the forefront of satisfiability research. Three of the most memorable issues were:

1. Some researchers lamented over the laser-like focus of many on the SAT competitions; they felt that not enough attention is being paid to the long-term scientific goal of understanding of why SAT is solvable in practice. Others argued in response that the SAT competitions are fun and community-building; they help motivate people to do worthy work with good intentions.
2. Related to the subject of competitions, a few researchers objected to their format, again based on scientific disagreement. There is still a rift between those designing “classical” CDCL-based SAT solvers, and those who use machine learning techniques to design algorithm “portfolios” selecting such SAT solvers to run on instances, and the SAT competition has developed rules to isolate the latter group from the rest of the solver submissions. The question of whether re-designing the competition in this way will positively (or negatively) influence further research is intriguing; it certainly was not resolved by this panel discussion.
3. Related to the subject of understanding SAT, there was extensive speculation by many parties on why SAT solvers tend to work so well in practice. Some pointed to the variable choice heuristics of solvers; some pointed to the clause learning of solvers; some posited that there must be inherent structure in most real-world SAT instances. Some asked (controversially) *if and why we should expect be able to understand SAT solvers at all*: SAT code and SAT instances solved in practice are so complex that perhaps humans simply cannot know, or cannot rigorously explain why practical SAT instances are solved so efficiently.

All in all, the thought-provoking discussion highlighted the diversity of attitudes and ideas that people bring to SAT research.

6 Examples of Outcomes of the Workshop

It is still a bit too early for any concrete publications to have resulted from the seminar, but participants have reported that the following papers, in different stages of preparation, were significantly influenced by discussions during the seminar:

- Albert Atserias, Massimo Lauria, and Jakob Nordström. **Narrow Proofs May Be Maximally Long**. Journal version in submissions, 2015.
- Armin Biere and Andreas Fröhlich. **Evaluating CDCL Variable Scoring Schemes**. To appear in *Proceedings of SAT’15*, September 2015.

- Armin Biere and Andreas Fröhlich. **SAT Solving and Stock Market Analysis**. Manuscript in preparation, 2015.
- Oliver Kullmann and João Marques-Silva. **Computing maximal autarkies with few and simple oracle queries**. To appear in *Proceedings of SAT'15*, September 2015.
- Massimo Lauria and Jakob Nordström. **Tight Size-Degree Bounds for Sums-of-Squares Proofs**. In *Proceedings of CCC'15*, June 2015.
- Jia Hui Liang, Vijay Ganesh, Ed Zulkoski, Atulan Zaman, Krzysztof Czarnecki. **Understanding VSIDS Branching Heuristics in Conflict-Driven Clause-Learning SAT Solvers**. Manuscript in submission, 2015.
- Jakob Nordström. **On the Interplay Between Proof Complexity and SAT Solving**. *ACM SIGLOG News*, July 2015.
- Mladen Mikša and Jakob Nordström. **A Generalized Method for Proving Polynomial Calculus Degree Lower Bounds**. In *Proceedings of CCC'15*, June 2015.

Making the connection to the panel discussion which we report on in Section 5, the Dagstuhl seminar week played an important role in stimulating a research project focused on a comprehensive empirical study to better understand the impact on performance of different features in modern CDCL SAT solvers. In joint work, Laurent Simon, João Marques-Silva, and Karem Sakallah have collected all non-random benchmarks from all SAT competitions and races (2002 to 2014) and instrumented both Minisat and Glucose to enable and disable their various options in order to pinpoint the effect of each option or combination of options on performance. The plan is to make this data available on a public website and provide extensive analysis of the data in a paper that is currently under preparation.

Other participants of the seminar have reported about at least six concrete research projects that resulted to a large part from contacts during the week at Dagstuhl. Since many of these projects are still in a start-up phase it would seem slightly premature to list concrete participants, but it can be mentioned that these projects involve researchers from INESC-ID Lisboa, Johannes Kepler University, KTH Royal Institute of Technology, Microsoft Research, Princeton University, RWTH Aachen, Swansea University, Universitat Politècnica de Catalunya, and University of Washington in various constellations. Several of these projects involves interdisciplinary research with both applied and theoretical components, and many seminar participants mentioned explicitly that the mix of theoreticians and practitioners at the seminar played a decisive role in making this happen.

7 Evaluation by Participants

In addition to the traditional Dagstuhl evaluation after the seminar, the organizing committee also arranged for a separate evaluation which specific questions about different aspects of the seminar. Below follows a summary of the answers – the full results are available at <http://www.csc.kth.se/~jakobn/dagstuhl15171/evaluation.php>.

In the post-seminar survey, the participants identified two major aspects of the seminar they enjoyed most: the networking opportunities between theoreticians and practitioners that the environment of Dagstuhl provided, and the high quality of the tutorial talks selected by the organizers. Many reported that they learned a substantial amount from the seminar talks.

However, the seminar was not immune from some negative feedback. Some found the tutorials too elementary, and felt there was not enough focus on talks with new results. Some felt that there should have been talks on the applications and general impact of SAT in

science and engineering. Some participants felt there was not enough proof complexity and others felt there was too much. A few did not like that some of the schedule extended into the late evening (which was the case for the Monday evening open problem session and the Wednesday evening panel discussion).

The seminar participants were polled before the seminar about some different aspects of the planning, and based on the results of this poll it was decided to have an open problem session on the first day. In the post-seminar survey, this decision was viewed favourably: 48% of respondents felt it was “definitely the right decision” and 40% felt it was “probably” the right decision. Some felt that the open problem session had too many problems, many of which were either too vague to fully grasp or too specific to be interesting; perhaps a “curated” open problem session would have been more effective.

Also based on results of the pre-seminar poll, we decided not to have a poster session, and an overwhelming majority felt this was the right decision in hindsight as well. Nevertheless, some did wish that there had been more opportunities to recreate “what happens at a poster session”: structured informal discussions about SAT research among many participants.

In general, much of the feedback contained the sentiment that more time for “guided” extended discussions among the entire group would have been useful. This is interesting when placed in the context of the feedback on the panel discussion (which was an intentionally guided discussion of SAT issues). Only slightly more than half of the respondents to the post-seminar survey felt that the panel was either “definitely” or “probably” a good idea with hindsight. Some enjoyed the panel, but others did not find the discussion fruitful. One participant, noting the abundance of experts at the seminar, suggested that a “town hall style” meeting (where everyone had the same chance to state their views) might have fared better.

All in all, the feedback from the participants was overwhelmingly positive. Many called the experience “great” or “fantastic” and thought the seminar had been “superbly organized” with “outstanding” talks. One participant even wrote that “[t]his was hands down the best Dagstuhl I have ever attended, and I have attended 10 so far”, and another respondent noted that “I and other people remarked that it seemed we could easily continue into a second week – people were refreshed rather than exhausted by the end of the seminar.” Many participants look forward to returning to Dagstuhl: in the post-seminar evaluation, 72% said they would definitely come again if invited to a similar seminar, and 20% said they would probably come again.

Participants

- Erika Abraham
RWTH Aachen, DE
- Albert Atserias
UPC – Barcelona, ES
- Gilles Audemard
CNRS – Lens, FR
- Paul Beame
University of Washington –
Seattle, US
- Christoph Berkholz
KTH Royal Institute of
Technology, SE
- Olaf Beyersdorff
University of Leeds, GB
- Armin Biere
Universität Linz, AT
- Nikolaj S. Bjørner
Microsoft Corporation –
Redmond, US
- Ilario Bonacina
University of Rome “La
Sapienza”, IT
- Sam Buss
University of California –
San Diego, US
- Amin Coja-Oghlan
Goethe-Universität – Frankfurt a.
M., DE
- Jessica Davies
IST Austria –
Klosterneuburg, AT
- Holger Dell
Universität des Saarlandes, DE
- Jan Elffers
KTH Royal Institute of
Technology, SE
- John Franco
University of Cincinnati, US
- Andreas Fröhlich
Universität Linz, AT
- Vijay Ganesh
University of Waterloo, CA
- Alexandra Goultiaeva
Google Waterloo, CA
- Martin Grohe
RWTH Aachen, DE
- Daniel Große
Universität Bremen, DE
- Edward A. Hirsch
Steklov Institute – St.
Petersburg, RU
- Holger H. Hoos
University of British Columbia –
Vancouver, CA
- Matti Järvisalo
University of Helsinki, FI
- Mikoláš Janota
INESC-ID – Lisboa, PT
- Manuel Kauers
Universität Linz, AT
- Oliver Kullmann
Swansea University, GB
- Massimo Lauria
KTH Royal Institute of
Technology, SE
- Daniel Le Berre
CNRS – Lens, FR
- Sharad Malik
Princeton University, US
- João Marques-Silva
INESC-ID – Lisboa, PT
- Nina Narodytska
Carnegie Mellon University, US
- Jakob Nordström
KTH Royal Institute of
Technology, SE
- Yakau Novikau
OneSpin Solutions –
München, DE
- Albert Oliveras
UPC – Barcelona, ES
- Pavel Pudlák
Academy of Sciences –
Prague, CZ
- Alexander Razborov
University of Chicago, US
- Kristin Yvonne Rozier
University of Cincinnati, US
- Karem A. Sakallah
University of Michigan – Ann
Arbor, US
- Martina Seidl
Universität Linz, AT
- Laurent Simon
University of Bordeaux, FR
- Niklas Sörensson
Mentor Graphics – Göteborg, SE
- Stefan Szeider
TU Wien, AT
- Alasdair Urquhart
University of Toronto, CA
- Allen Van Gelder
University of California – Santa
Cruz, US
- Marc Vinyals
KTH Royal Institute of
Technology, SE
- Magnus Wahlström
Royal Holloway University of
London, GB
- Ryan Williams
Stanford University, US



Challenges and Trends in Probabilistic Programming

Edited by

Gilles Barthe¹, Andrew D. Gordon², Joost-Pieter Katoen³, and Annabelle McIver⁴

- 1 IMDEA Software – Madrid, ES, gjbarthe@gmail.com
- 2 Microsoft Research UK – Cambridge, GB, adg@microsoft.com
- 3 RWTH Aachen University, DE, katoen@cs.rwth-aachen.de
- 4 Macquarie University – Sydney, AU, annabelle.mciver@mq.edu.au

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15181 “Challenges and Trends in Probabilistic Programming”. Probabilistic programming is at the heart of machine learning for describing distribution functions; Bayesian inference is pivotal in their analysis. Probabilistic programs are used in security for describing both cryptographic constructions (such as randomised encryption) and security experiments. In addition, probabilistic models are an active research topic in quantitative information now. Quantum programs are inherently probabilistic due to the random outcomes of quantum measurements. Finally, there is a rapidly growing interest in program analysis of probabilistic programs, whether it be using model checking, theorem proving, static analysis, or similar. Dagstuhl Seminar 15181 brought researchers from these various research communities together so as to exploit synergies and realize cross-fertilisation.

Seminar April 27–30, 2015 – <http://www.dagstuhl.de/15181>

1998 ACM Subject Classification D.1 Programming Techniques, D.2.4 Software/Program Verification, D.3 Programming Languages, D.4.6 Security and Protection, F. Theory of Computation, G.1.6 Optimization, G.3 Probability and Statistics, I.2.2 Automatic Programming, I.2.4 Knowledge Representation Formalisms and Methods

Keywords and phrases Bayesian networks, differential privacy, machine learning, probabilistic programs, security, semantics, static analysis, verification

Digital Object Identifier 10.4230/DagRep.5.4.123

Edited in cooperation with Benjamin Kaminski

1 Executive Summary

Gilles Barthe

Andrew D. Gordon

Joost-Pieter Katoen

Annabelle McIver

License  Creative Commons BY 3.0 Unported license
© Gilles Barthe, Andrew D. Gordon, Joost-Pieter Katoen, and Annabelle McIver

Probabilistic programming languages

Probabilistic programs are programs, written in languages like C, Java, LISP, or ML, with two added constructs: (1) the ability to draw values at random from probability distributions, and (2) the ability to condition values of variables in a program through observations. A variety of probabilistic programming languages have been defined such as Church, Infer.NET,



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Challenges and Trends in Probabilistic Programming, *Dagstuhl Reports*, Vol. 5, Issue 4, pp. 123–141

Editors: Gilles Barthe, Andrew D. Gordon, Joost-Pieter Katoen, and Annabelle McIver



DAGSTUHL
REPORTS

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

and IBAL. Church is based on the Lisp model of the lambda calculus, containing pure Lisp as its deterministic subset, whereas Infer.NET is a Microsoft developed language akin to C# and compiles probabilistic programs into inference code¹. Probabilistic programs can be used for modelling complex phenomena from biology and social sciences. By doing so, we get the benefits of programming languages (rigorous semantics, execution, testing and verification) to these problem domains. More than a decade ago, McIver and Morgan defined a probabilistic programming language in the style of Dijkstra's guarded command language, referred to as pGCL. Besides the usual language constructs in Dijkstra's GCL such as non-deterministic choice, it features a probabilistic choice where the probability distribution may be parametric. For instance, the assignment $x += 1 [p]$ skip increments the variable x by one with probability p , and keeps the value of x unchanged with probability $1-p$, where p is an unknown real value from the range $[0, 1]$. Quantum programming languages such as qGCL and a quantum extension of C++ are also related, as their operational semantics is typically a probabilistic model so as to model the effect of measurements on the quantum state.

The importance of probabilistic programming

The applications of probabilistic programs mainly lie in four domains: (1) machine learning, (2) security, (3) randomised algorithms, and – though to a somewhat lesser extent – (4) quantum computing. Whereas the application in the field of randomised algorithms is evident, let us briefly describe the importance for the other three fields.

Machine learning

A Bayesian generative model consists of a prior distribution over some parameters, together with a sampling distribution (or likelihood) that predicts outputs of the model given its inputs and parameters. Bayesian inference in machine learning consists of training such a model to infer the posterior distribution of the parameters and hence to make predictions. In the probabilistic programming approach to Bayesian inference, the user simply writes the prior and sampling distributions as probabilistic programs, and relies on a compiler to generate code to perform inference and make predictions. Such compilers often operate by considering the program as defining a probabilistic graphical model. Graphical models were pioneered by Judea Pearl and others, and are extensively described in the comprehensive text by Koller and Friedman (2009). They are widely used in statistics and machine learning, with diverse application areas including speech recognition, computer vision, biology, and reliability analysis. Probabilistic graphical models allow specification of dependences between random variables via generative models, as well as conditioning of random variables using phenomena or data observed in the real world. A variety of inference algorithms have been developed to analyse and query such models, e.g., Gibbs sampling methods, Metropolis-Hastings and belief propagation. The probabilistic programming approach has seen growing interest within machine learning over the last 10 years and it is believed – see <http://probabilistic-programming.org/wiki/Home> – that this approach within AI has the potential to fundamentally change the way that community understands, designs, builds, tests and deploys probabilistic systems.

¹ For academic use, it is free to use: <http://research.microsoft.com/infernet>.

Security

Ever since Goldwasser and Micali – recipients of the ACM Turing Award in 2013 – introduced probabilistic encryption, probability has played a central role in cryptography: virtually all cryptographic algorithms are randomized, and have probabilistic security guarantees. Similarly, perturbing outputs with probabilistic noise is a standard tool for achieving privacy in computations; for instance, differential privacy achieves privacy-preserving data-mining using probabilistic noise. Cryptographic algorithms and differentially private algorithms are implemented as probabilistic programs; more singularly, one common approach for reasoning about these algorithms is using the code-based game-based approach, proposed by Bellare and Rogaway, in which not only the algorithms, but also their security properties and the hardness properties upon which their security relies, are expressed as probabilistic programs, and can be verified using (a relational variant of) Hoare logic. This code-based approach is key to recent developments in verified cryptography. Quantitative information flow is another important field in security where probabilistic programs and models play an important role. Here, the key question is to obtain quantitative statements about the leakage of certain information from a given program.

Quantum computing

Quantum programs are used to describe quantum algorithms and typically are quantum extensions of classical while-programs. Whereas in classical computation, we use a type to denote the domain of a variable, in quantum computation, a type is the state space of a quantum system denoted by some quantum variable. The state space of a quantum variable is the Hilbert space denoted by its type. According to a basic postulate of quantum mechanics, the unique way to acquire information about a quantum system is to measure it. Therefore, the essential ingredient in a quantum program is the ability to perform measurements of quantum registers, i.e., finite sequences of distinct quantum variables. The state space of a quantum register is the tensor product of the state spaces of its quantum variables. In executing the statement `measure M[q]; S`, quantum measurement M will first be performed on quantum register q , and then a sub-program S will be selected to be executed next according to the outcome of the measurement. The essential difference between a measurement statement and a classical conditional statement is that the state of program variables is changed after performing the measurement. As the outcome of a measurement is probabilistic, quantum programs are thus inherently probabilistic.

Program analysis

On the other hand, there is a recent rapidly growing trend in research on probabilistic programs which is more in line with traditional programming languages. This focuses on aspects such as efficient compilation, static analysis, program transformations, and program verification. To mention a few, Cousot *et al.* recently extended the framework of abstract interpretation to probabilistic programs (2012), Gordon *et al.* introduced `Tabular`, a new probabilistic programming language (2014), Di Pierro *et al.* apply probabilistic static analysis (2010), Rajamani, Gordon *et al.* have used symbolic execution to perform Bayesian reasoning on probabilistic programs with loops (2013), Katoen, McIver *et al.* have developed invariant synthesis technique for linear probabilistic programs (2010), and Geldenhuys *et al.* considered probabilistic symbolic execution (2012).

Achievements of this seminar

The objective of the seminar was to bring together researchers from four separate (but related) communities to learn from each other, with the expectation that a better understanding between these communities would open up new opportunities for research and collaboration.

Participants attending the seminar represented all four themes of the original proposal: machine learning, quantitative security, (probabilistic) program analysis and quantum computing. The programme consisted of both tutorials and presentations on any topic within these themes. The tutorials provided a common ground for discussion, and the presentations gave insight into the current state of an area, and summarised the challenges that still remain. The tutorial topics were determined by consulting the participants prior to the seminar by means of a questionnaire.

Although the programme was primarily constructed around the tutorials and standard-length presentations (each around 30 minutes), the organisers made sure that time was always available for short, impromptu talks (sometimes of only 5 minutes) where participants were able to outline a relevant challenge problem or to draw attention to a new research direction or connection that had become apparent during the meeting.

This open forum for exploring links between the communities has led to the following specific achievements:

1. An increased understanding between the disciplines, especially between program verification and probabilistic programming.
2. A demonstration that the mathematical models for reasoning about machine learning algorithms and quantitative security are very similar, but that their objectives are very different. This close relationship at a foundational level suggests theoretical methods to tackle the important challenge of understanding privacy in a data mining context.
3. Evidence that probabilistic programming, analysis and verification of probabilistic programs, can have a broad impact in the design of emerging infrastructures, such as software-defined networks.

The feedback by the participants was very positive, and it was encouraged to organise a workshop or similar event in the future to foster the communication between the different communities, in particular between program verification and probabilistic programming.

We were aware of many new conversations between researchers inspired by the formal talks as well as the mealtime discussions. Already at least one paper (see below) with content inspired by the meeting is accepted for publication, and we are aware of several other new lines of work.

Acknowledgement

The organisers thank Benjamin Kaminski for his support in compiling this report and in several organisational issues.

References

- 1 A. Ścibior, Z. Ghahramani and A. D. Gordon. *Practical Probabilistic Programming with Monads*. ACM SIGPLAN Haskell Symposium 2015, Vancouver, Canada, 3–4 September 2015.

2 Table of Contents

Executive Summary

Gilles Barthe, Andrew D. Gordon, Joost-Pieter Katoen, and Annabelle McIver . . . 123

Overview of Talks

Proving Differential Privacy in Hoare Logic <i>Gilles Barthe</i>	129
Reasoning about Approximate and Uncertain Computation <i>Michael Carbin</i>	129
Equivalence of (Higher-Order) Probabilistic Programs <i>Ugo Dal Lago</i>	129
A Topological Quantum Calculus <i>Alessandra Di Pierro</i>	130
Dyna: A Circuit Programming Language for Statistical AI <i>Jason Eisner</i>	131
Probabilistic Termination <i>Luis Maria Ferrer Fioriti</i>	131
Computability of conditioning: approximate inference and conditional independence <i>Cameron Freer</i>	132
Tabular: A Schema-Driven Probabilistic Programming Language <i>Andrew D. Gordon</i>	132
Conditioning in Probabilistic Programming <i>Friedrich Gretz</i>	133
On the Hardness of Almost-Sure Termination <i>Benjamin Kaminski</i>	133
Distinguishing Hidden Markov Chains <i>Stefan Kiefer</i>	133
Tutorial on Probabilistic Programming Languages <i>Angelika Kimmig</i>	134
Rational Protection against Timing Attacks <i>Boris Köpf</i>	134
Probabilistic Programming for Security <i>Piotr Mardziel</i>	135
Three Tokens Suffice <i>Joel Ouaknine</i>	135
The Design and Implementation of Figaro <i>Avi Pfeffer</i>	136
Dual Abstractions of Hidden Markov Models: A Monty Hell Puzzle <i>Tahiry Rabehaja</i>	136
Types and Modules for Probabilistic Programming Languages <i>Norman Ramsey</i>	137

Conditioning by Lazy Partial Evaluation <i>Chung-chieh Shan</i>	137
NetKAT – A Formal System for the Verification of Networks <i>Alexandra Silva</i>	137
WOLFE: Practical Machine Learning Using Probabilistic Programming and Optimization <i>Sameer Singh</i>	138
Recent Results in Quantitative Information Flow <i>Geoffrey Smith</i>	138
Tutorial on Probabilistic Programming in Machine Learning <i>Frank Wood</i>	139
Quantum Programming: From Superposition of Data to Superposition of Programs <i>Mingsheng Ying</i>	139
Counterexample-Guided Polynomial Quantitative Loop Invariants by Lagrange Interpolation <i>Lijun Zhang</i>	140
Participants	141

3 Overview of Talks

3.1 Proving Differential Privacy in Hoare Logic

Gilles Barthe (IMDEA Software, Spain)

License © Creative Commons BY 3.0 Unported license
© Gilles Barthe

Main reference Gilles Barthe, Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, César Kunz, Pierre-Yves Strub: Proving Differential Privacy in Hoare Logic. CSF 2014

Differential privacy is a rigorous privacy policy which provides individuals strong guarantees in the context of privacy-preserving data mining. Thanks to its rigorous definition, differential privacy is amenable to formal verification. Using a notion of (ϵ, δ) -lifting which generalizes the standard definition of lifting used in probabilistic process algebra, we develop a relational program logic to prove that probabilistic computations are differentially private.

3.2 Reasoning about Approximate and Uncertain Computation

Michael Carbin (Microsoft Research – Redmond, US)

License © Creative Commons BY 3.0 Unported license
© Michael Carbin

Joint work of Carbin, Michael; Misailovic, Sasa; Rinard, Martin

Main reference M. Carbin, S. Misailovic, M. C. Rinard, “Verifying Quantitative Reliability for Programs that Execute on Unreliable Hardware,” in Proc. of 28th ACM SIGPLAN Conf. on Object-Oriented Programming, Systems, Languages and Applications (OOPSLA/SPLASH’13), pp. 33–52, ACM, 2013.

URL <http://dx.doi.org/10.1145/2509136.2509546>

Many modern applications implement large-scale computations (e.g., machine learning, big data analytics, and financial analysis) in which there is a natural trade-off between the quality of the results that the computation produces and the performance and cost of executing the computation.

Exploiting this fact, researchers have recently developed a variety of new mechanisms that automatically change the structure and execution of an application to enable it to meet its performance requirements. Examples of these mechanisms include skipping portions of the application’s computation and executing the application on fast and/or energy-efficient unreliable hardware systems whose operations may silently produce incorrect results.

I present a program verification and analysis system, Rely, whose novel verification approach makes it possible to verify the safety, security, and accuracy of the approximate applications that these mechanisms produce. Rely also provides a program analysis that makes it possible to verify the probability that an application executed on unreliable hardware produces the same result as if it were executed on reliable hardware.

3.3 Equivalence of (Higher-Order) Probabilistic Programs

Ugo Dal Lago (University of Bologna, IT)

License © Creative Commons BY 3.0 Unported license
© Ugo Dal Lago

We introduce program equivalence in the context of higher-order probabilistic functional programs. The canonical notion of equivalence, namely context equivalence, has the nice

property of prescribing equivalent programs to behave the same in any context, but has the obvious drawback of being based on a universal quantification over all contexts. We show how the problem can be overcome by going through a variation of Abramsky's applicative bisimulation. We finally hints at the role of equivalence in cryptographic proof.

3.4 A Topological Quantum Calculus

Alessandra Di Pierro (University of Verona, IT)

License  Creative Commons BY 3.0 Unported license
© Alessandra Di Pierro

The work by Richard Feynman in the 1980s, and by Seth Lloyd and many others starting in the 1990s showed how a wide range of realistic quantum systems can be simulated by using quantum circuits, i.e. a quantum computer. In 1989, Edward Witten established a connection between problem solving and quantum field theories; he discovered a strong analogy between the Jones polynomial (an important knot invariant in topology) and Topological Quantum Field Theory (TQFT). Some years later, this discovery inspired a new form of quantum computation, called Topological Quantum Computation (TQC). A topological quantum computer would be computationally as powerful as a standard one. Nevertheless, Witten's discovery of the connection between TQFT and the value of the Jones polynomial at particular roots of unity implicitly suggested an efficient quantum algorithm for the approximation of the Jones polynomial, a problem which classically belongs to the P# complexity class and for which the standard quantum computing algorithmic techniques currently known do not provide any speed-up. Topological Quantum Computation is based on the existence of two-dimensional particles called anyons, whose statistics substantially differ from what we can observe in a three-dimensional quantum system. The behaviour of anyons can be described via the statistics observed after exchanging one particle with another. This exchange rotates the system's quantum state and produces non trivial phases. The idea of using such systems for computing is due to Alexei Kitaev and dates back to 1997. Since then, TQC has been mainly studied in the realm of physics and mathematics, while only recently the algorithmic and complexity aspects of this computational paradigm has been investigated in the area of computer science. Following this line, in this work we revisit TQC from the perspective of computability theory and investigate the question of computational universality for TQC, namely the definition of a anyonic quantum computer that is able to simulate any program on any other anyonic quantum computer. To this aim we introduce a formal calculus for TQC whose definition uses a language which is neither physical nor categorical but rather logical (if we look at the calculus as an equational theory) or programming-oriented (by considering it as an abstract model of computation). We adopt a formalism similar to the lambda-calculus that we call anyonic lambda-calculus. This calculus is essentially a re-writing system consisting of two transformation rules, namely variable substitution (as in the classical lambda-calculus) and a second one representing the braiding of anyons. The function definition scheme is exactly the same as Church's lambda-calculus. However, differently from the latter, the anyonic lambda-calculus represents an anyonic computer, that is a quantum system of anyons where computation occurs by braiding a fixed number of anyons among them for some fixed time. This is an approximation process that allows us to achieve approximate results, i.e. results that are not exact although their precision can be fixed arbitrarily high. For this calculus we provide an operational semantics in the form

of a rewriting system and we show a property of confluence which takes into account the approximate nature of topological quantum computation.

References

- 1 Pachos, J. K., *Introduction to Topological Quantum Computation*, Cambridge U.P., 2012.
- 2 Wang, Z., *Topological Quantum Computation*, American Mathematical Soc., 2010.
- 3 Witten, E., *Topological quantum field theory*, Commun. Math. Phys **117** (1988), pp. 353–386.
- 4 Aharonov, D., V. Jones and Z. Landau, *A polynomial quantum algorithm for approximating the jones polynomial*, in: Proceedings of STOC’06 (2006), pp. 427–436.
- 5 Barendregt, H. P., *The Lambda Calculus*, Studies in Logic and the Foundations of Mathematics **103**, North-Holland, 1991, revised edition.
- 6 Freedman, M. H., A. Kitaev, M. J. Larsen and Z. Wang, *Topological Quantum Computation*, Physical Review Letters **40** (2001), p. 120402.
- 7 Jones, V. F. R., *A polynomial invariant for knots via Von Neumann algebras*, Bull. American Mathematical Society (New Series) **12** (1985), pp. 103–111.
- 8 Kitaev, A., *Fault-tolerant quantum computation by anyons*, Annals of Physics **303** (1997).

3.5 Dyna: A Circuit Programming Language for Statistical AI

Jason Eisner (Johns Hopkins University, US)

License  Creative Commons BY 3.0 Unported license
© Jason Eisner

The Dyna programming language is intended to provide an declarative abstraction layer for building systems in ML and AI. A Dyna program specifies a generalized circuit that defines named quantities from other named quantities, using weighted Horn clauses with aggregation. The Dyna runtime must efficiently find a fixpoint of this circuit and maintain it under changes to the inputs. The language is an extension of logic programming with non-boolean values, evaluation, aggregation, types, and modularity. We illustrate how Dyna supports design patterns in AI, allowing extremely concise specifications of various algorithms, and we discuss the implementation decisions that are left to the system. Finally, we also sketch a preliminary design for P-Dyna, a probabilistic modeling language that can be embedded within Dyna and is based on augmenting Dyna’s circuits with randomness.

3.6 Probabilistic Termination

Luis Maria Ferrer Fioriti (Universität des Saarlandes, DE)

License  Creative Commons BY 3.0 Unported license
© Luis Maria Ferrer Fioriti

Joint work of Ferrer Fioriti, Luis María; Hermanns, Holger

Main reference L. M. Ferrer Fioriti, H. Hermanns, “Probabilistic Termination: Soundness, Completeness, and Compositionality,” in Proc. of the 42nd Annual ACM SIGPLAN-SIGACT Symp. on Principles of Programming Languages (POPL’15), pp. 489–501, ACM, 2015.

URL <http://dx.doi.org/10.1145/2676726.2677001>

We propose a framework to prove almost sure termination for probabilistic programs with real valued variables. It is based on ranking supermartingales, a notion analogous to ranking functions on nonprobabilistic programs. The framework is proven sound and complete for a

meaningful class of programs involving randomization and bounded nondeterminism. We complement this foundational insight by a practical proof methodology, based on sound conditions that enable compositional reasoning and are amenable to a direct implementation using modern theorem provers. This is integrated in a small dependent type system, to overcome the problem that lexicographic ranking functions fail when combined with randomization. Among others, this compositional methodology enables the verification of probabilistic programs outside the complete class that admits ranking supermartingales.

3.7 Computability of conditioning: approximate inference and conditional independence

Cameron Freer (MIT – Cambridge, US)

License © Creative Commons BY 3.0 Unported license
© Cameron Freer

Joint work of Ackerman, Nathanael; Avigad, Jeremy; Freer, Cameron; Roy, Daniel; Rute, Jason

Main reference N. L. Ackerman, C. E. Freer, D. M. Roy, “On the computability of conditional probability,” arXiv:1005.3014v2 [math.LO], 2011.

URL <http://arxiv.org/abs/1005.3014v2>

Here we address three key questions at the theoretical and algorithmic foundations of probabilistic programming – and probabilistic modeling more generally – that can be answered using tools from probability theory, computability and complexity theory, and non-parametric Bayesian statistics. Which Bayesian inference problems can be automated, and which cannot? Can probabilistic programming languages represent the stochastic processes at the core of state-of-the-art nonparametric Bayesian models? And if not, can we construct useful approximations?

3.8 Tabular: A Schema-Driven Probabilistic Programming Language

Andrew D. Gordon (Microsoft Research UK – Cambridge, GB)

License © Creative Commons BY 3.0 Unported license
© Andrew D. Gordon

Joint work of Gordon, Andrew D.; Graepel, Thore; Rolland, Nicolas; Russo, Claudio; Borgstrom, Johannes; John Guiver, John

Main reference A. D. Gordon, T. Graepel, N. Rolland, C. Russo, J. Borgstrom, J. Guiver, “Tabular: A Schema-driven Probabilistic Programming Language,” in Proc. of the 41st ACM SIGPLAN-SIGACT Symp. on Principles of Programming Languages (POPL’14), pp. 321-334, ACM, 2014.

URL <http://dx.doi.org/10.1145/2535838.2535850>

We propose a new kind of probabilistic programming language for machine learning. We write programs simply by annotating existing relational schemas with probabilistic model expressions. We describe a detailed design of our language, Tabular, complete with formal semantics and type system. A rich series of examples illustrates the expressiveness of Tabular. We report an implementation, and show evidence of the succinctness of our notation relative to current best practice. Finally, we describe and verify a transformation of Tabular schemas so as to predict missing values in a concrete database. The ability to query for missing values provides a uniform interface to a wide variety of tasks, including classification, clustering, recommendation, and ranking.

3.9 Conditioning in Probabilistic Programming

Friedrich Gretz (RWTH Aachen, DE)

License © Creative Commons BY 3.0 Unported license
© Friedrich Gretz

Joint work of Gretz, Friedrich; Jansen, Nils; Kaminski, Benjamin Lucien; Katoen, Joost-Pieter; McIver, Annabelle; Olmedo, Federico

Main reference F. Gretz, N. Jansen, B.L. Kaminski, J.-P. Katoen, A. McIver, F. Olmedo, “Conditioning in Probabilistic Programming,” arXiv:1504.00198v1 [cs.PL], 2015.

URL <http://arxiv.org/abs/1504.00198v1>

In practical applications of probabilistic programming such as machine learning often the goal is to infer parameters of a probabilistic model from observed data. The used inference methods are entirely based on sampling and statistical methods. At the same time probabilistic programs in the realm of formal methods have a formal semantics that precisely captures the distribution generated by a program. First formal analysis techniques for such programs are emerging. Thus the question is if we can bring together two areas and apply formal methods to machine learning. Our work goes in this direction by introducing observations in a minimalistic core probabilistic language called pGCL. We are able to extend two existing equivalent semantics to conditional probability distributions. Our semantics are sound even for programs that do not necessarily terminate with probability one. We explain how non-determinism in the model can be handled in the operational semantics and why it is problematic for denotational semantics. We conclude with applications of our semantics. For one, we show how in principle we can reason about properties of probabilistic programs with observations. Second, we show how our semantics enable us to formally proof the validity of program transformations which are useful in practise.

3.10 On the Hardness of Almost-Sure Termination

Benjamin Kaminski (RWTH Aachen, DE)

License © Creative Commons BY 3.0 Unported license
© Benjamin Kaminski

Joint work of Kaminski, Benjamin; Katoen, Joost-Pieter

We study the computational hardness of computing expected outcomes and deciding (universal) (positive) almost-sure termination of probabilistic programs. It is shown that computing lower and upper bounds of expected outcomes is Σ_1^0 - and Σ_2^0 -complete, respectively. Deciding (universal) almost-sure termination as well as deciding whether the expected outcome of a program equals a given rational value is shown to be Π_2^0 -complete. Finally, it is shown that deciding (universal) positive almost-sure termination is Σ_2^0 -complete (Π_3^0 -complete).

3.11 Distinguishing Hidden Markov Chains

Stefan Kiefer (University of Oxford, UK)

License © Creative Commons BY 3.0 Unported license
© Stefan Kiefer

Hidden Markov Chains (HMCs) are commonly used mathematical models of probabilistic systems. They are specified by a Markov Chain, capturing the probabilistic behavior of a

system, and an output function specifying the outputs generated from each of its states. One of the important problems associated with HMCs is the problem of identification of the source of outputs generated by one of a number of known HMCs. We report on progress on this problem.

3.12 Tutorial on Probabilistic Programming Languages

Angelika Kimmig (*KU Leuven, BE*)

License  Creative Commons BY 3.0 Unported license
© Angelika Kimmig

Probabilistic programming languages combine programming languages with probabilistic primitives as well as general purpose probabilistic inference techniques. They thus facilitate constructing and querying complex probabilistic models. This tutorial provides a gentle introduction to the field through a number of core probabilistic programming concepts. It focuses on probabilistic logic programming (PLP), but also connects to related areas such as statistical relational learning and probabilistic databases. The tutorial illustrates the concepts through examples, discusses the key ideas underlying inference in PLP, and touches upon parameter learning, language extensions, and applications in areas such as bioinformatics, object tracking and information processing.

An interactive tutorial can be found at <https://dtai.cs.kuleuven.be/problog/>.

References

- 1 Luc De Raedt and Angelika Kimmig. *Probabilistic (logic) programming concepts*. Machine Learning, 2015. <http://dx.doi.org/10.1007/s10994-015-5494-z>

3.13 Rational Protection against Timing Attacks

Boris Köpf (*IMDEA Software – Madrid, ES*)

License  Creative Commons BY 3.0 Unported license
© Boris Köpf

Joint work of Doychev, Goran; Köpf, Boris

Main reference G. Doychev, B. Köpf, “Rational Protection against Timing Attacks,” in Proc. of the IEEE 28th Computer Security Foundations Symp. (CSF’15), pp. 526–536, IEEE, 2015; revised version available from author’s webpage.

URL <http://dx.doi.org/10.1109/CSF.2015.39>

URL <http://software.imdea.org/~bkoepf/papers/csf15.pdf>

Timing attacks can effectively recover keys from cryptosystems. While they can be defeated using constant-time implementations, this defensive approach comes at the price of a performance penalty. One is hence faced with the problem of striking a balance between performance and security against timing attacks.

This talk presents a game-theoretic approach to the problem, for the case of cryptosystems based on discrete logarithms. Namely, we identify the optimal countermeasure configuration as an equilibrium in a game between a resource-bounded timing adversary who strives to maximize the probability of key recovery, and a defender who strives to reduce the cost while maintaining a certain degree of security. The key novelty in our approach are bounds for the probability of key recovery, which are expressed as a function of the countermeasure configuration and the attack strategy of the adversary.

We put our techniques to work for a library implementation of ElGamal. A highlight of our results is that we can formally justify the use of an aggressively tuned but (slightly) leaky implementation over a defensive constant-time implementation, for some parameter ranges. The talk concludes with an outlook on how static analysis, probabilistic programming, and machine learning can help with performing similar analyses for more general classes of programs.

References

- 1 Goran Doychev and Boris Köpf. *Rational Protection against Timing Attacks*. 28th IEEE Computer Security Foundations Symposium (CSF). 2015

3.14 Probabilistic Programming for Security

Piotr Mardziel (University of Maryland – College Park, US)

License  Creative Commons BY 3.0 Unported license
© Piotr Mardziel

Probabilistic inference is a powerful tool for reasoning about hidden data from restricted observations and probabilistic programming is a convenient means of expressing and mechanizing this process. Likewise the same approaches can be used to model adversaries learning about secrets. Security, however, often relies on formal guarantees not typical in machine learning applications. In this talk we will compare and contrast the two applications of probabilistic programming and present our work on approximate probabilistic inference that is sound relative to quantitative measures of information security.

3.15 Three Tokens Suffice

Joel Ouaknine (University of Oxford, GB)

License  Creative Commons BY 3.0 Unported license
© Joel Ouaknine

Herman's self-stabilisation algorithm, introduced 25 years ago, is a well-studied synchronous randomised protocol for enabling a ring of N processes collectively holding any odd number of tokens to reach a stable state in which a single token remains. Determining the worst-case expected time to stabilisation is the central outstanding open problem about this protocol. It is known that there is a constant h such that any initial configuration has expected stabilisation time at most hN^2 . Ten years ago, McIver and Morgan established a lower bound of $4/27 \approx 0.148$ for h , achieved with three equally-spaced tokens, and conjectured this to be the optimal value of h . A series of papers over the last decade gradually reduced the upper bound on h , with the present record (achieved last year) standing at approximately 0.156. In a paper currently under review, we prove McIver and Morgan's conjecture and establish that $h = 4/27$ is indeed optimal.

In the talk, I would like to describe Herman's protocol, consider examples, discuss related work and some of the history of the problem, and present a very brief schematic overview of the approach.

3.16 The Design and Implementation of Figaro

Avi Pfeffer (Charles River Analytics – Cambridge, US)

License  Creative Commons BY 3.0 Unported license
© Avi Pfeffer

In this talk, I present some of the motivations for the design of the Figaro probabilistic programming system (PPS) and describe the approach to implementing the system, particularly in regards to factored inference algorithms. Figaro is a PPS that is able to represent a very wide range of probabilistic models and provides automated inference algorithms for reasoning with those models. Figaro is designed to be easy to integrate with applications and data and to support many modeling frameworks, like functional and object-oriented paradigms, directed and undirected models, hybrid models with discrete and continuous variables, and dynamic models. Figaro is designed as an embedded library in Scala; you write Scala programs to construct and operate on Figaro models. This provides numerous advantages such as support for integration and the ability to construct models programmatically. Figaro has been applied to a number of applications in areas like cyber security, climate prediction, and system health monitoring.

Many PPSs use sampling algorithms such as Markov chain Monte Carlo for inference and Figaro also provides such algorithms. However, in Figaro, we are trying to make factored inference algorithms like variable elimination and belief propagation viable for probabilistic programming. These algorithms are often the best performing for graphical models, but they can be difficult to apply to probabilistic programs because they assume a factor graph of fixed, finite structure. We address this problem with two main ideas. First, lazy factored inference partially expands a model to a finite depth and bounds the influence of the unexpanded part of the model on the query, thereby enabling factored algorithms to be applied even when the factor graph is very large or infinite. We have shown the ability to produce bounds on problems where sampling and other factored algorithms cannot operate. Second, structured inference uses the model definition to automatically decompose a difficult factor graph into subproblems. Each of these subproblems can be solved using a different solver. We have shown that using different algorithms on different subproblems can yield a significant improvement in accuracy without incurring additional computation cost.

3.17 Dual Abstractions of Hidden Markov Models: A Monty Hell Puzzle

Tahiry Rabehaja (Macquarie University – Sydney, AU)

License  Creative Commons BY 3.0 Unported license
© Tahiry Rabehaja

Hidden Markov Models, HMMs, are mathematical models of Markov processes whose state is hidden but from which information can leak via channels. They are typically represented as 3-way joint probability distributions. We use HMMs as denotations of probabilistic hidden state sequential programs, after recasting them as “abstract” HMMs, computations in the Giry monad, and equipping them with a partial order of increasing security. We then present uncertainty measures as a generalisation of the extant diversity of probabilistic entropies, and we propose characteristic analytic properties for them. Based on that, we give a “backwards”, uncertainty-transformer semantics for HMMs, dual to the “forwards” abstract

HMMs. The backward semantics is specifically aimed towards a source level reasoning method for probabilistic hidden state sequential programs. [Joint work with Annabelle McIver and Carroll Morgan.]

I will be talking about channels, Markov processes and HMMs through a small Monty Hell puzzle. We will see that they are pieces of the single unified framework of abstract HMMs which in turn admit backward interpretations as UM-transformers. The transformer semantics constitutes the logical basis towards a source level quantitative analysis of programs with hidden states.

3.18 Types and Modules for Probabilistic Programming Languages

Norman Ramsey (Tufts University – Medford, US)

License  Creative Commons BY 3.0 Unported license
© Norman Ramsey

Many probabilistic programming languages include only core-language constructs for deterministic computation, plus primitives for probabilistic modeling and inference. We hypothesize that, like many other special-purpose languages, probabilistic languages could benefit from linguistic apparatus that has been found to be helpful in more general settings – in particular, types and modules. To support this hypothesis, we introduce the model, which resembles an ML module, but which in addition to a type part and a value part, also enjoys a distribution part. These parts are described in a model type, which is analogous to an ML signature or interface. In both a model and its type, distribution part is described compositionally by a collection of bindings to random variables. To explore the values of these ideas, we present a family of model types, at different levels of abstraction, and a corresponding model, of a problem in seismic detection (provided by Stuart Russell). Many challenges remain, of which the most pressing may be specifying the desire to learn a predictive posterior distribution.

3.19 Conditioning by Lazy Partial Evaluation

Chung-chieh Shan (Indiana University – Bloomington, US)

License  Creative Commons BY 3.0 Unported license
© Chung-chieh Shan

We review how to define measures mathematically, express them as programs, and run them as samplers. We then show how to define conditioning mathematically and implement it as a program transformation.

3.20 NetKAT – A Formal System for the Verification of Networks

Alexandra Silva (Radboud University Nijmegen, NL)

License  Creative Commons BY 3.0 Unported license
© Alexandra Silva

This talk will describe NetKAT, a formal system to program and verify networks. I will describe work from the following two articles:

1. Carolyn Jane Anderson, Nate Foster, Arjun Guha, Jean-Baptiste Jeannin, Dexter Kozen, Cole Schlesinger, and David Walker, NetKAT: Semantic Foundations for Networks. POPL 14.
2. Nate Foster, Dexter Kozen, Matthew Milano, Alexandra Silva, and Laure Thompson. A Coalgebraic Decision Procedure for NetKAT. POPL 15.

3.21 WOLFE: Practical Machine Learning Using Probabilistic Programming and Optimization

Sameer Singh (University of Washington – Seattle, US)

License © Creative Commons BY 3.0 Unported license
© Sameer Singh

Performing machine learning with existing toolkits on large datasets is quite a frustrating experience: each toolkit focuses on its own subclass of machine learning techniques, have their own different interface of how much of the underlying system is surfaced to the user, and don't support the iterative development that is required to tune machine learning algorithms and achieve satisfactory predictors.

In this talk we present Wolfe, a declarative machine learning stack consisting of three crucial components: (1) Language: a math-like syntax embedded in Scala to concisely specify arbitrarily complex machine learning systems that unify most existing, and future, techniques, (2) Interpreter that transforms the declarative description into efficient code that scales to large-datasets, and (3) REPL: A new iPython-like IDE for Scala that supports the unique features for machine learning such as visualizing structured data, probability distributions, and state of optimization.

3.22 Recent Results in Quantitative Information Flow

Geoffrey Smith (Florida International University – Miami, US)

License © Creative Commons BY 3.0 Unported license
© Geoffrey Smith

Main reference G. Smith, "Recent Developments in Quantitative Information Flow (Invited Tutorial)," in Proc. of the 30th Annual ACM/IEEE Symp. on Logic in Computer Science (LICS'15), pp. 23–31, IEEE, 2015; pre-print available from author's webpage.

URL <http://dx.doi.org/10.1109/LICS.2015.13>

URL <http://users.cis.fiu.edu/~smithg/papers/lics15.pdf>

In computer security, it is frequently necessary in practice to accept some leakage of confidential information. This motivates the development of theories of Quantitative Information Flow aimed at showing that some leaks are "small" and therefore tolerable. We describe the fundamental view of channels as mappings from prior distributions on secrets to hyperdistributions, which are distributions on posterior distributions, and we show how g-leakage provides a rich family of operationally-significant measures of leakage. We also discuss two approaches to achieving robust judgments about leakage: notions of capacity and a robust leakage ordering called composition refinement.

3.23 Tutorial on Probabilistic Programming in Machine Learning

Frank Wood (University of Oxford, GB)

License © Creative Commons BY 3.0 Unported license
© Frank Wood

This tutorial covers aspects of probabilistic programming that are of particular importance in machine learning in a way that is meant to be accessible and interesting to programming languages researchers. Example programs and inference are demonstrated in the Anglican programming language and examples of new inference algorithms applicable to inference in probabilistic programming systems, in particular the particle cascade, are provided.

3.24 Quantum Programming: From Superposition of Data to Superposition of Programs

Mingsheng Ying (University of Technology – Sydney, AU)

License © Creative Commons BY 3.0 Unported license
© Mingsheng Ying

We extract a novel quantum programming paradigm – superposition of programs – from the design idea of a popular class of quantum algorithms, namely quantum walk-based algorithms. The generality of this paradigm is guaranteed by the universality of quantum walks as a computational model.

A new quantum programming language QGCL is then proposed to support the paradigm of superposition of programs. This language can be seen as a quantum extension of Dijkstra’s GCL (Guarded Command Language). Alternation (case statement) in GCL splits into two different notions in the quantum setting: classical alternation (of quantum programs) and quantum alternation, with the latter being introduced in QGCL for the first time. Quantum alternation is the key program construct for realizing the paradigm of superposition of programs.

The denotational semantics of QGCL are defined by introducing a new mathematical tool called the guarded composition of operator-valued functions. Then the weakest precondition semantics of QGCL can straightforwardly be derived.

Another very useful program construct in realizing the quantum programming paradigm of superposition of programs, called quantum choice, can be easily defined in terms of quantum alternation. The relation between quantum choices and probabilistic choices is clarified through defining the notion of local variables.

Furthermore, quantum recursion with quantum control flow is defined based on second quantisation method.

We believe that this new quantum programming paradigm can help to further exploit the unique power of quantum computing.

3.25 Counterexample-Guided Polynomial Quantitative Loop Invariants by Lagrange Interpolation

Lijun Zhang (Chinese Academy of Sciences, CN)

License  Creative Commons BY 3.0 Unported license
© Lijun Zhang

We apply multivariate Lagrange interpolation to synthesizing polynomial quantitative loop invariants for probabilistic programs. We reduce the computation of an quantitative loop invariant to solving constraints over program variables and unknown coefficients. Lagrange interpolation allows us to find constraints with less unknown coefficients. Counterexample-guided refinement furthermore generates linear constraints that pinpoint the desired quantitative invariants. We evaluate our technique by several case studies with polynomial quantitative loop invariants in the experiments.

Participants

- Christel Baier
TU Dresden, DE
- Gilles Barthe
IMDEA Software – Madrid, ES
- Johannes Borgström
Uppsala University, SE
- Michael Carbin
Microsoft Res. – Redmond, US
- Aleksandar Chakarov
Univ. of Colorado – Boulder, US
- Ugo Dal Lago
University of Bologna, IT
- Alessandra Di Pierro
University of Verona, IT
- Jason Eisner
Johns Hopkins University –
Baltimore, US
- Yuan Feng
University of Technology –
Sydney, AU
- Luis Maria Ferrer Fioriti
Universität des Saarlandes, DE
- Cédric Fournet
Microsoft Research UK –
Cambridge, GB
- Cameron Freer
MIT – Cambridge, US
- Marco Gaboardi
University of Dundee, GB
- Andrew D. Gordon
Microsoft Research UK –
Cambridge, GB
- Friedrich Gretz
RWTH Aachen, DE
- Johannes Hölzl
TU München, DE
- Chung-Kil Hur
Seoul National University, KR
- Benjamin Kaminski
RWTH Aachen, DE
- Joost-Pieter Katoen
RWTH Aachen, DE
- Stefan Kiefer
University of Oxford, GB
- Angelika Kimmig
KU Leuven, BE
- Boris Köpf
IMDEA Software – Madrid, ES
- Pasquale Malacaria
Queen Mary University of
London, GB
- Vikash Mansinghka
MIT – Cambridge, US
- Piotr Mardziel
University of Maryland –
College Park, US
- Annabelle McIver
Macquarie Univ. – Sydney, AU
- Joel Ouaknine
University of Oxford, GB
- Catuscia Palamidessi
INRIA Saclay –
Île-de-France, FR
- David Parker
University of Birmingham, GB
- Avi Pfeffer
Charles River Analytics –
Cambridge, US
- Tahiry Rabehaja
Macquarie Univ. – Sydney, AU
- Sriram K. Rajamani
Microsoft Research India –
Bangalore, IN
- Norman Ramsey
Tufts University – Medford, US
- Chung-chieh Shan
Indiana University –
Bloomington, US
- Alexandra Silva
Radboud Univ. Nijmegen, NL
- Sameer Singh
University of Washington –
Seattle, US
- Geoffrey Smith
Florida International University –
Miami, US
- Andreas Stuhlmüller
MIT Cambridge & Stanford
University, US
- Frank Wood
University of Oxford, GB
- Mingsheng Ying
University of Technology –
Sydney, AU
- Lijun Zhang
Chinese Academy of Sciences, CN



Qualification of Formal Methods Tools

Edited by

Darren Cofer¹, Gerwin Klein², Konrad Slind³, and Virginie Wiels⁴

1 Rockwell Collins – Minneapolis, US, darren.cofer@rockwellcollins.com

2 NICTA – Sydney, AU, gerwin.klein@nicta.com.au

3 Rockwell Collins – Minneapolis, US, konrad.slind@rockwellcollins.com

4 ONERA – Toulouse, FR, virginie.wiels@onera.fr

Abstract

Formal methods tools have been shown to be effective at finding defects in and verifying the correctness of safety-critical systems, many of which require some form of certification. However, there are still many issues that must be addressed before formal verification tools can be used as part of the certification of safety-critical systems. For example, most developers of avionics systems are unfamiliar with which formal methods tools are most appropriate for different problem domains. Different levels of expertise are necessary to use these tools effectively and correctly. In most certification processes, a tool used to meet process objectives must be *qualified*. The qualification of formal verification tools will likely pose unique challenges.

Seminar April 26–29, 2015 – <http://www.dagstuhl.de/15182>

1998 ACM Subject Classification D.2.4 Software/program verification, F.3.1 Specifying and Verifying and Reasoning about Programs, G.4 Mathematical Software

Keywords and phrases Dependable systems, Certification, Qualification, Formal methods, Verification tools

Digital Object Identifier 10.4230/DagRep.5.4.142

1 Executive Summary

Darren Cofer

Gerwin Klein

Konrad Slind

Virginie Wiels

License  Creative Commons BY 3.0 Unported license
© Darren Cofer, Gerwin Klein, Konrad Slind, and Virginie Wiels

Motivation and objectives

Dagstuhl Seminar 13051, *Software Certification: Methods and Tools*, convened experts from a variety of software-intensive domains (automotive, aircraft, medical, nuclear, and rail) to discuss software certification challenges, best practices, and the latest advances in certification technologies. One of the key challenges identified in that seminar was tool qualification. Tool qualification is the process by which certification credit may be claimed for the use of a software tool. The purpose of tool qualification is to provide sufficient confidence in the tool functionality so that its output may be trusted. Tool qualification is, therefore, a significant aspect of any certification effort. Seminar participants identified a number of needs in the area of formal methods tool qualification. Dagstuhl Seminar 15182 *Qualification of Formal Methods Tools*, was organized to address these needs.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Qualification of Formal Methods Tools, *Dagstuhl Reports*, Vol. 5, Issue 4, pp. 142–159

Editors: Darren Cofer, Gerwin Klein, Konrad Slind, and Virginie Wiels



DAGSTUHL
REPORTS

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

Software tools are used in development processes to automate life cycle activities that are complex and error-prone if performed by humans. The use of such tools should, in principle, be encouraged from a certification perspective to provide confidence in the correctness of the software product. Therefore, we should avoid unnecessary barriers to tool qualification which may inadvertently reduce the use of tools that would otherwise enhance software quality and confidence.

Most software tools are not used in isolation, but are used as part of a complex tool chain requiring significant integration effort. In general, these tools have been produced by different organizations. We need to develop better and more reliable methods for integrating tools from different vendors (including university tools, open source tools, and commercial tools).

A given software tool may be used in different application domains having very different requirements for both certification and tool qualification. Furthermore, the methods and standards for tool development varies across domains. Consistent qualification requirements across different domains would simplify the process.

Despite the additional guidance provided for the avionics domain in recently published standards (DO-178C, DO-330, and DO-333), there are still many questions to be addressed. For one thing, most practicing engineers are unaware of how to apply different categories of formal verification tools. Even within a particular category, there are a wide variety of tools, often based on fundamentally different approaches, each with its own strengths and weaknesses.

If formal verification is used to satisfy DO-178C objectives, DO-333 requires the applicant to provide evidence that the underlying method is sound, i.e., that it will never assert something is true when it is actually false, allowing application software errors to be missed that should have been detected. Providing an argument for the soundness of a formal verification method is highly dependent on the underlying algorithm on which the method is based. A method may be perfectly sound when used one way on a particular type of problem and inherently unsound when used in a different way or on a different type of problem. While these issues may be well understood in the research community, they are not typically collected in one place where a practitioner can easily find them. It is also not realistic to expect avionics developers to be able to construct an argument for the soundness of a formal method without help from experts in the field.

At the same time, it is also important to not make the cost of qualification of formal methods tools so great as to discourage their use. While it is tempting to hold formal verification tools to a higher standard than other software tools, making their qualification unnecessarily expensive could do more harm than good.

The objectives of this Dagstuhl Seminar were to

- investigate the sorts of assurances that are necessary and appropriate to justify the application of formal methods tools throughout all phases of design in real safety-critical settings,
- discuss practical examples of how to qualify different types of formal verification tools, and
- explore promising new approaches for the qualification of formal methods tools for the avionics domain, as well as in other domains.

Accomplishments

Qualification is not a widely understood concept outside of those industries requiring certification for high-assurance, and different terminology is used in different domains. The seminar was first a way of sharing knowledge from certification experts so that formal methods researchers could better understand the challenges and barriers to the use of formal methods tools.

The seminar also included presentations from researchers who have developed initial approaches to address qualification requirements for different classes of formal methods tools. We were especially interested in sharing case studies that are beginning to address tool qualification challenges. These case studies include tools based on different formal methods (model checking, theorem proving, abstract interpretation).

As a practical matter, we focussed much of our discussion on the aerospace domain since there are published standards addressing both formal methods and tool qualification for avionics software. The seminar also included researchers from other domains (nuclear, railway) so we could better understand the challenges and tool qualification approaches that are being discussed in those domains.

We managed to bridge a lot of the language between the certification domains, mostly railway, avionics, and nuclear, and bits of automotive, and related the qualification requirements to each other. Some of the otherwise maybe less stringent schemes (e.g. automotive) can end up having stronger qualification requirements, because formal methods are not specifically addressed in them. There is some hope that DO-333 might influence those domains, or be picked up by them in the future, to increase the use of FM tools which would increase the quality of systems.

For the academic tool provider side, we worked out and got the message across that tool qualification can be a lot easier and simpler than what we might strive for academically, and discussed specific tools in some detail, clarifying what would be necessary for a concrete qualification. Finally, we also investigated tool architectures that make tools easier to qualify (verification vs code generation).

2 Table of Contents

Executive Summary

Darren Cofer, Gerwin Klein, Konrad Slind, and Virginie Wiels 142

Overview of Talks

Please check my 500K LOC of Isabelle
June Andronick 146

Compiling avionics software with the CompCert formally verified compiler
Sandrine Blazy 148

Qualification of Formal Methods Tools and Tool Qualification with Formal Methods
Matteo Bordin 148

Are You Qualified for This Position? An Introduction to Tool Qualification
Darren Cofer 148

Sharing experience on SAT-based formal verification toolchain qualification in the railway domain
Rémi Delmas 149

Qualification of PVS for Systematic Design Verification of a Nuclear Shutdown System
Mark Lawford 150

How much is CompCert's proof worth, qualification-wise?
Xavier Leroy 151

Certificates for the Qualification of the Model Checker Kind 2
Alain Mebsout 153

Towards Certification of Network Calculus
Stephan Merz 153

Tool Qualification Strategy for Abstract Interpretation-based Static Analysis Tools
Markus Pister 154

Tool Qualification in the Railway Domain
Werner Schuetz 154

FM Tool Trust Propositions
Konrad Slind 155

DO-330 Tool Qualification: An experience report
Lucas Wagner 156

Discussion Groups

Why qualify a formal methods tool? 156

How to qualify a formal methods tool? 157

Compiler qualification strategies 157

Comparison of qualification in different domains 158

Participants 159

3 Overview of Talks

3.1 Please check my 500K LOC of Isabelle

June Andronick (UNSW – Sydney, AU)

License  Creative Commons BY 3.0 Unported license
© June Andronick

The seL4 microkernel has been formally proved correct [2], from binary code, up to high level requirements, using the Isabelle theorem prover [5]. In this talk we first gave an overview of seL4 development and proof guarantees and assumptions. We then explored what would be needed for a (hypothetical) certification of seL4 according to DO-178 (the software certification standard for airborne systems on commercial aircraft [6]), including a potential qualification of Isabelle according to DO-330 (tool qualification guidelines [7]).

The seL4 microkernel is a small operating system kernel, of roughly 10,000 lines of C code, designed to be a high-performance, secure, safe, and reliable foundation for a wide variety of application domains. It provides isolation and controlled communication to applications running on top of it, allowing trusted applications to run alongside untrusted, legacy code such as a whole Linux instance.

seL4 is the world’s most verified kernel [2], with a full functional correctness proof, showing that the binary code is a correct implementation of the high-level functional specification, plus security proofs, showing that seL4 enforces integrity and confidentiality. All the proofs have been conducted in the Isabelle/HOL theorem prover, apart from the binary-to-C correctness proof, which uses some SMT solvers and HOL4 models and proofs. The combined Isabelle proofs amount to about 500,000 lines of Isabelle models and proof scripts.

For this Dagstuhl seminar of tool qualification, we have put ourselves in the situation of wanting to certify seL4 for use in an avionics context, and therefore needing to qualify the tools used in its formal verification, here mainly Isabelle, according to DO-330. Following the discussions and presentations from the seminar, we investigated the following question:

What would be needed to qualify Isabelle, for the objective of using the proof of functional correctness of seL4 to justify that the code is complete and correct with respect to its high-level specification?

From our understanding of the qualification process, we propose to answer the following questions.

1. Justify that the *method (Interactive Theorem Proving) is suitable:*

Since the property we are showing is functional correctness, it requires a high-level of expressiveness to precisely model the code and specification; such high level of expressiveness implies a loss of decidability, and therefore requires user’s input to perform the proof. Interactive theorem proving fits precisely with those requirements. To justify this to a certifier, we could refer to peer-reviewed papers or point to examples of projects using interactive theorem provers to prove functional correctness.

2. Justify that the *method (Isabelle-style deduction) is sound:*

Isabelle’s logic is based on a very small kernel that needs to be trusted: a dozen axioms, that have been manually validated. All extensions are derived from first principles and checked by this kernel. The only ways of adding axioms is through (conservative) definitions and through explicit axioms and tracked oracles (e.g. sorried lemmas). To justify this to a certifier, we could again refer to peer-reviewed papers, the *HOL-report* [1], or the formally verified HOL-light [3] and CakeML implementations [4].

3. Justify that the *tool* (Isabelle) correctly implements the method:

This would require us to show that only the standard distribution theory HOL is used, that no *axiom* commands are used after the theory HOL, that no “sorry” and “cheat_tac” commands are used, and other technical corner-cases that should be documented. When these conditions are met, only true theorems in HOL can be derived. Evidences for this question would ideally be a small verified proof checker for Isabelle (using e.g. cakeML and providing efficient proof terms).

4. Justify the *correct use of the tool* (Isabelle):

This would consist in checking that the above conditions (no axioms, no sorries, etc) are satisfied in the specific example of the proof under consideration. This is where the title of this talk comes from.

5. Justify that the tool (Isabelle) is helping *meeting the objective*:

This would require showing that the model of C used is a correct representation of C, that the model of the specification is a correct representation of the expected behavior, and that the formalisation of the property (here refinement) is a correct representation of the objective (here that the code is complete and correct with respect to its high-level specification). The seL4 verification includes high-level security proofs, which aim at justifying that the specification satisfies the expected behaviors. Evidence for the C model and refinement statement could be done by review, inspection and testing. As a community, it would also be helpful to provide documentation and training material on how to *read* formal specification, to allow certifiers and non-experts to convince themselves that the statements and properties make sense. Then they only need to trust the experts and peer-reviewed papers that the proof script will indeed provide an evidence that the statement is true, that the property is satisfied.

Acknowledgments

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

- 1 Mike Gordon. HOL: A machine oriented formulation of higher-order logic. Technical report, University of Cambridge Computer Laboratory, 1985.
- 2 Gerwin Klein, June Andronick, Kevin Elphinstone, Toby Murray, Thomas Sewell, Rafal Kolanski, and Gernot Heiser. Comprehensive formal verification of an OS microkernel. *ACM Transactions on Computer Systems*, 32(1):2:1–2:70, February 2014.
- 3 Ramana Kumar, Rob Arthan, Magnus O. Myreen, and Scott Owens. HOL with definitions: Semantics, soundness, and a verified implementation. In Gerwin Klein and Ruben Gamboa, editors, *Interactive Theorem Proving (ITP)*, pages 308–324. Springer, 2014.
- 4 Ramana Kumar, Magnus Myreen, Michael Norrish, and Scott Owens. CakeML: A verified implementation of ML. In Peter Sewell, editor, *ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 179–191, San Diego, jan 2014. ACM Press.
- 5 Tobias Nipkow, Lawrence Paulson, and Markus Wenzel. *Isabelle/HOL – A Proof Assistant for Higher-Order Logic*, volume 2283 of *Lecture Notes in Computer Science*. Springer, 2002.
- 6 RTCA. *DO-178C, Software Considerations in Airborne Systems and Equipment Certification*.
- 7 RTCA. *DO-330, Software Tool Qualification Considerations*.

3.2 Compiling avionics software with the CompCert formally verified compiler

Sandrine Blazy (IRISA – Rennes, FR)

License  Creative Commons BY 3.0 Unported license
© Sandrine Blazy

Compilers are complicated pieces of software that sometimes contain bugs causing wrong executable code to be silently generated from correct source programs. In turn, this possibility of compiler-introduced bugs diminishes the assurance that can be obtained by applying formal methods to source code.

This talk gives an overview of the CompCert project: an ongoing experiment in developing and formally proving correct a realistic, moderately-optimizing compiler from a large subset of C to PowerPC, ARM and x86 assembly languages. The correctness proof, mechanized using the Coq proof assistant, establishes that the generated assembly code behaves exactly as prescribed by the semantic of the C source, eliminating all possibilities of compiler-introduced bugs and generating unprecedented confidence in this compiler.

3.3 Qualification of Formal Methods Tools and Tool Qualification with Formal Methods

Matteo Bordin (AdaCore – Paris, FR)

License  Creative Commons BY 3.0 Unported license
© Matteo Bordin

This work focuses on the return of experience in the relation between Formal Methods and Tool Qualification. We explored two main application domains: the qualification of formal methods tools and the use of formal methods for tool qualification. In the first case, we present our work in qualifying an abstract interpretation tool (CodePeer) and a formal verification tool (SPARK) in a DO-178 context. In the second case, we focus instead on a lightweight use of formal methods to help the qualification of an automated code generator from Simulink models. This second experience is particularly interesting as it describes how we used Ada 2012 contracts (pre/post-condition) to formally describe in first-order logic the behavior of a code generator. Such specification is not used to statically verify the code generator, but rather as a run-time oracle that checks that the tool executes accordingly to its specifications. Differently from other similar experiences, and quite to our surprise, we realized that the specification in the form of pre/post-conditions significantly differed from the implementation algorithm.

3.4 Are You Qualified for This Position? An Introduction to Tool Qualification

Darren Cofer (Rockwell-Collins – Minneapolis, US)

License  Creative Commons BY 3.0 Unported license
© Darren Cofer

Formal methods tools have been shown to be effective at finding defects in and verifying the correctness of safety-critical systems such as avionics systems. The recent release of

DO-178C and the accompanying Formal Methods supplement DO-333 will make it easier for developers of software for commercial aircraft to obtain certification credit for the use of formal methods.

However, there are still many issues that must be addressed before formal verification tools can be injected into the design process for safety-critical systems. For example, most developers of avionics systems are unfamiliar with which formal methods tools are most appropriate for different problem domains. Different levels of expertise are necessary to use these tools effectively and correctly. Evidence must be provided of a formal method's soundness, a concept that is not well understood by most practicing engineers. Finally, DO-178C requires that a tool used to meet its objectives must be qualified in accordance with the tool qualification document DO-330. The qualification of formal verification tools will likely pose unique challenges.

Qualification is not a widely understood concept outside of those industries requiring certification for high-assurance, and different terminology is used in different domains. This talk provided an overview of certification and qualification requirements for the civil aviation domain so that formal methods researchers can better understand the challenges and barriers to the use of formal methods tools. Topics covered included a summary of certification processes and objectives for avionics software, requirements for qualification of tools used in software development and verification, and how formal methods tools fit into the certification environment.

3.5 Sharing experience on SAT-based formal verification toolchain qualification in the railway domain

Rémi Delmas (ONERA – Toulouse, FR)

License  Creative Commons BY 3.0 Unported license
© Rémi Delmas

The goal of the talk is to fuel the reflexion and discussion about formal verification tool qualification in the aerospace domain according to the new DO-333 guidelines, by sharing previous experience on tool qualification in the railway domain under CENELEC SIL-* requirements. The talk describes a formal verification toolchain based on SAT solvers and k-induction used in the railway domain for the verification of safety properties of interlocking and communication-based train control systems. The tool in question has been used to earn certification credits, by replacing tests with formal properties verification, in real world railway control systems. In particular, the talk describes how the tool chain's architecture, development and V&V process was designed in order to meet CENELEC SIL-4 tool qualification requirements, using implementation diversification, semantic equivalence checking, proof-logging/proof-checking. The talk also highlights the various non-technical issues that surround formal verification tool qualification, which nevertheless must be taken into account to ensure the success of formal verification in industrial applications.

3.6 Qualification of PVS for Systematic Design Verification of a Nuclear Shutdown System

Mark Lawford (McMaster University – Hamilton, CA)

License  Creative Commons BY 3.0 Unported license
© Mark Lawford

The Systematic Design Verification (SDV) process used on the redesign of the Darlington Nuclear Generating Station originated in the difficulties encountered in receiving regulatory approval for Canada’s first computer based reactor shutdown system (SDS) [4]. The SDV process for the redesign project made use of tabular expressions for the Software Requirements Specification (SRS) and the Software Design Description (SDD). Completeness and consistency of the tabular expressions and the conformance of the SDD to the SRS were established using the automated theorem prover PVS [3]. The process used to qualify PVS for use in this context is described below and related to the latest version of IEC 61508.

The qualification required the use of manual proof to mitigate against potential undetected errors that might be caused by a failure of PVS, i.e., all of the proofs performed in the PVS theorem prover also had to be done by hand. The standard IEC 61508 (2nd ed) in part 4 provides a classification of tools according to whether they are *software on-line support tools* that can directly influence system safety at run time, or *software off-line support tools* that support a phase of the software development lifecycle and that cannot directly influence the safety-related system during its run time. Software off-line support tools are further broken down into three subclasses:

- T1:** generates no outputs which can directly or indirectly contribute to the executable code (including data) of the safety related system; (e.g. a text editor, a requirements or design support tool with no automatic code generation capabilities, configuration control tools)
- T2:** supports the test or verification of the design or executable code, where errors in the tool can fail to reveal defects but cannot directly create errors in the executable software; (e.g. a test harness generator, test coverage measurement tool, static analysis tool)
- T3:** generates outputs which can directly or indirectly contribute to the executable code of the safety related system (e.g., an optimising compiler where the relationship between the source code program and the generated object code is not obvious, a compiler that incorporates an executable run-time package into the executable code).

According to this classification, PVS as used on the Darlington Redesign Project would be a T2 tool since it is being used to verify a design and a tool failure could fail to reveal an error but not introduce an error into the executable.

In IEC 61508-3 (2nd ed) it states that:

7.4.4.5 An assessment shall be carried out for offline support tools in classes T2 and T3 to determine the level of reliance placed on the tools, and the potential failure mechanisms of the tools that may affect the executable software. Where such failure mechanisms are identified, appropriate mitigation measures shall be taken.

Since a failure mechanism is that PVS has a bug that causes a proof to succeed when it should have failed, we needed a mitigation strategy. The strategy chosen was to redo all proofs manually. Although this mitigation strategy might appear to defeat much of the benefit of using a formal methods tool, PVS could still be used to quickly check design iterations and the manual checks only needed to be performed on the final work product to mitigate PVS’s failure modes. Still, the final manual proofs were tedious and required significant effort.

A proposal is made for a revised Tabular Expression Toolbox that makes use of PVS and an SMT solver to eliminate the need for manual review in order to gain tool qualification. A prototype implementation of the Tabular Expression Toolbox is described in [1].

References

- 1 Eles, C. and Lawford, M. (2011). A tabular expression toolbox for matlab/simulink. In *3rd NASA Formal Methods Symposium*, volume 6617 of *LNCS*, pages 494–499. Springer-Verlag.
- 2 Pang, L., Wang, C.-W., Lawford, M., and Wassyng, A. (2014). Formalizing and verifying function blocks using tabular expressions and pvs. In *Second International Workshop on Formal Techniques for Safety-Critical Systems (FTSCS 2013)*, volume 419 of *Communications in Computer and Information Science*, pages 163–178. Springer.
- 3 Wassyng, A. and Lawford, M. (2003). Lessons learned from a successful implementation of formal methods in an industrial project. In Araki, K., Gnesi, S., and Mandrioli, D., editors, *FME 2003: International Symposium of Formal Methods Europe Proceedings*, volume 2805 of *Lecture Notes in Computer Science*, pages 133–153, Pisa, Italy. Springer-Verlag.
- 4 Wassyng, A., Lawford, M. S., and Maibaum, T. S. (2011). Software certification experience in the canadian nuclear industry: lessons for the future. In *Proceedings of the ninth ACM international conference on Embedded software*, EMSOFT '11, pages 219–226, New York, NY, USA. ACM.

3.7 How much is CompCert’s proof worth, qualification-wise?

Xavier Leroy (INRIA – Le Chesnay, FR)

License © Creative Commons BY 3.0 Unported license
© Xavier Leroy

Intuitively as well as experimentally (cf. the Csmith compiler testing project), the formal verification of the CompCert C compiler generates much confidence that it is free of miscompilation issues. How can we derive certification credit from this formal verification, in the context of a DO-330 / DO-333 tool qualification? This question is being investigated within the Verasco project (ANR-11-INSE-03; <http://verasco.imag.fr/>).

Consider first the formally-verified part of the CompCert C compiler. This part goes from abstract syntax for the CompCert subset of C to abstract syntax for the assembly language of the target processor. This part contains all the optimizations and almost all code generation algorithms. For this part, we see a plausible mapping between parts of the Coq development and DO-330 concepts:

- The “specifications” part of the Coq development constitutes most of the (high-level) tool requirements. This part comprises the abstract syntax and operational semantics of the CompCert C and CompCert assembly languages, as well as the high-level statement of compiler correctness, namely preservation of semantics during compilation, with preservation of properties as a corollary.
- The “code” part of the Coq development map to the low-level tool requirements. This part comprises all compilation algorithms (written in pure functional, executable style in Coq’s specification language) as well as the abstract syntaxes of the intermediate languages used. It is comparable to the pseudocode or Simulink/Scade models that are used as low-level requirements in other certifications.
- The “proof” part of the Coq development automates the verification activities between the (high-level) tool requirements and the low-level tool requirements. This part contains the proofs of semantic preservation for every compilation pass, the proofs of semantic

soundness for every static analysis, as well as the operational semantics for the intermediate languages.

A first difficulty is that the “specifications”, “code” and “proof” parts are not clearly separated in CompCert’s Coq development, owing to good mathematical style (theorems and their proofs come just after definitions) and also to the use of dependently-typed data structures. It would be useful to develop a “slicing” tool for Coq that extracts the various parts of the development by tracing dependencies.

The source code for the compiler, in DO-330 parlance, corresponds to the OCaml code that is generated from the “code” part of the Coq development by Coq’s extraction facility. The executable compiler, then, is obtained by OCaml compilation. Here, we are in familiar territory: automatic code generation followed by compilation. However, suitable confidence arguments must be provided for Coq’s extraction and for OCaml’s compilation. Several approaches were discussed during the meeting, ranging from dissimilar implementations to Coq-based validation of individual runs of the executable compiler.

At the other end of the DO-330 sequence of refinements, we are left with the tool operational requirements, which have to be written in informal prose, with references to the ISO C 1999 language standard, the ISA reference manuals for the target architecture, and coding standards such as MISRA C. The verification activities here are essentially manual, and include for example relating the CompCert C formal semantics with the informal specifications in ISO C 1999 and MISRA. Such a relation can be built from appropriate tests, since CompCert provides a reference interpreter that provides an executable, testable form of its C formal semantics.

All in all, the formal proof of CompCert does not eliminate the need for manual verifications, but it reduces their scope tremendously: from manual verification of a full optimizing compiler to manual verification of formal semantics for C and assembly languages. For example, changes to the “code” part of the compiler (e.g. adding new optimizations, modifying the intermediate languages, etc) need no new manual verification activities, as long as the “specification” part of the compiler is unchanged.

To finish, we need to consider the parts of the CompCert C compiler that are not formally verified yet: uphill of the verified part, the transformations from C source text to CompCert C abstract syntax (preprocessing, tokenization, parsing, type-checking, pre-simplifications, production of an abstract syntax tree); downhill, the transformation from assembly abstract syntax to ELF executables (assembling and linking). CompCert provides an independent checker that validates a posteriori the assembling and linking phases. Likewise, some of the uphill passes were formally verified recently (parsing and type-checking). Nonetheless, many of the uphill passes lack formal specifications and therefore must be verified by conventional, test-based means.

In conclusions, the qualification of an optimizing compiler to the highest quality levels has never been attempted before, and might very well be too expensive to be worth the effort. A formal compiler verification such as CompCert’s has high potential to reduce these costs. However, much work remains to take full advantage of this potential.

3.8 Certificates for the Qualification of the Model Checker Kind 2

Alain Mebsout (University of Iowa – Iowa City, US)

License © Creative Commons BY 3.0 Unported license
© Alain Mebsout

Joint work of Mebsout, Alain; Tinelli, Cesare

This talk presents a technique for generating proof certificates in the model checker Kind 2 as an alternate path of qualification with respect to DO-178C. This is put in perspective with the qualification that was conducted for the SMT solver Alt-Ergo at Airbus for use in the development of the A350. Alt-Ergo was qualified wrt DO-178B as a backend solver for Caveat to verify C code of the pre-flight inspection. On the other hand, Kind 2 generates proof certificates which allows to shift the trust from the model checker to the proof checker (LFSC). Certificates for the actual model checking algorithm are generated as SMT2 files and verified by an external SMT solver. The translation from Lustre to the internal first-order logic representation is verified in a lightweight way by proving observational equivalence between independent frontends (for the moment JKind and Kind 2). This proof is actually carried by Kind 2 itself and generates in turn SMT2 certificates.

3.9 Towards Certification of Network Calculus

Stephan Merz (INRIA Nancy – Villers-lès-Nancy, FR)

License © Creative Commons BY 3.0 Unported license
© Stephan Merz

Joint work of Boyer, Marc; Fejoz, Loïc; Mabilie, Etienne; Merz, Stephan

Main reference E. Mabilie, M. Boyer, L. Fejoz, S. Merz, “Towards Certifying Network Calculus,” in Proc. of the 4th Int’l Conf. on Interactive Theorem Proving (ITP’13), LNCS, Vol. 7998, pp. 484–489, Springer, 2013.

URL http://dx.doi.org/10.1007/978-3-642-39634-2_37

Network Calculus (NC) is an established theory for determining bounds on message delays and for dimensioning buffers in the design of networks for embedded systems. It is supported by academic and industrial tool sets and has been widely used, including for the design and certification of the Airbus A380 AFDX backbone. However, tool sets used for developing certified systems need to be qualified, which requires substantial effort and makes them rigid, even when deficiencies are subsequently detected. Result checking may be a worthwhile complement, since the use of a qualified (and highly trustworthy) checker could replace qualifying the analysis tool itself. In this work, we experimented an encoding of the fundamental theory of NC in the interactive proof assistant Isabelle/HOL and used it to check the results of a prototypical NC analyzer.

3.10 Tool Qualification Strategy for Abstract Interpretation-based Static Analysis Tools

Markus Pister (AbsInt – Saarbrücken, DE)

License © Creative Commons BY 3.0 Unported license
© Markus Pister

Joint work of Kästner Daniel, Pister Markus, Gebhard Gernot, Ferdinand Christian
Main reference D. Kästner, M. Pister, G. Gebhard, C. Ferdinand, “Reliability of WCET Analysis,” Embedded Real Time Software and Systems Congress (ERTSS’14), Toulouse, France, 2014.

In automotive, railway, avionics and healthcare industries more and more functionality is implemented by embedded software. A failure of safety-critical software may cause high costs or even endanger human beings. Also for applications which are not highly safety-critical, a software failure may necessitate expensive updates.

Safety-critical software has to be certified according to the pertinent safety standard to get approved for release. Contemporary safety standards including DO-178C, IEC-61508, ISO-26262, and EN-50128 require the identification of potential functional and non-functional hazards and to demonstrate that the software does not violate the relevant safety goals. If tools are used to satisfy the corresponding verification objectives, an appropriate tool qualification is mandatory to show functional correctness of the tool behavior with respect to the operational context.

To ensure functional program properties, automatic or model-based testing and formal techniques like model checking are becoming more widely used. For non-functional properties identifying a safe end-of-test criterion is a hard problem since failures usually occur in corner cases and full test coverage cannot be achieved.

For some non-functional program properties this problem is solved by abstract interpretation-based static analysis techniques which provide full control and data coverage and yield provably correct results. Like model checking and theorem proving, abstract interpretation belongs to the formal software verification methods. AbsInt provides abstract interpretation-based static analyzers to determine safety-guarantees on the worst-case execution time (aiT) and stack consumption (StackAnalyzer) as well as to prove the absence of runtime errors (Astree) in safety-critical software.

This talk focuses on our tool qualification strategy of the above mentioned verification tools, which are increasingly adopted by industry in their validation activities for safety-critical software. First, we will give an overview of the tools and their role within the analyzed system’s certification process. We then outline the required activities for a successful tool qualification of our static analyzers alongside their correspondingly produced data.

3.11 Tool Qualification in the Railway Domain

Werner Schuetz (Thales – Wien, AT)

License © Creative Commons BY 3.0 Unported license
© Werner Schuetz

In this presentation we give an overview of the relevant standards applicable to the rail domain. EN50128 is concerned with software, while EN50129 addresses system issues.

This presentation focuses on tool qualification. The 2011 edition of EN50128 is the first to include requirements on “Support Tools and Languages”. To this end it defines three tool classes. T3 tools directly or indirectly produce code or data that is used in the safety-related

system. T2 tools are verification tools that may fail to detect an error but cannot introduce an error themselves. T1 tools do not contribute directly or indirectly to the executable code or data.

This presentation discusses the requirements on support tools and how they apply to the three tool classes. Comparison with the relevant aerospace standards (DO178C, DO330) is partly given.

In an appendix we briefly analyze which “Formal Methods” are contained in the 2011 edition of EN50128.

3.12 FM Tool Trust Propositions

Konrad Slind (Rockwell-Collins – Minneapolis, USA)

License  Creative Commons BY 3.0 Unported license
© Konrad Slind

An interactive theorem proving (ITP) system is a complex piece of software that bundles a great deal of functionality together. Beyond their core theorem proving task, which can employ highly complex algorithms, these systems provide extensibility, rich interfaces for users, interaction with host operating systems, etc. And yet, ITP systems are claimed to provide very high assurance. It is our purpose to take a close look at this state of affairs and explain the justifications for this claim.

We introduce the notion of the *trust proposition* to organize the discussion: it helps the consumer of a theorem prover’s output understand what the full assurance story is, by breaking the overall trust proposition down to subcomponents. In particular, we identify the work product of an ITP as a collection of theories, which formalize the artifact under scrutiny, plus properties and proofs. This work product can be trusted, provided the following conditions are met:

1. **Trusted Basis** The support theories are trusted;
2. **Trusted Extension** The newly introduced types, constants, definitions, and axioms are trusted;
3. **Valid Model** The support theories plus newly introduced types, constants, definitions, and axioms accurately model the artifact under scrutiny;
4. **Sound Logic** The proof system is sound;
5. **Correct Implementation** The proof system and extension mechanisms are correctly implemented
6. **Correct Libraries** The libraries used in the implementation are correctly implemented;
7. **Correct Compilation** The compiler correctly compiles the libraries and the implementation of the proof system;
8. **Correct Execution** The machine correctly runs the executable; and
9. **Trusted IO** The input and output of the ITP can be trusted.

3.13 DO-330 Tool Qualification: An experience report

Lucas Wagner (Rockwell Collins – Cedar Rapids, US)

License  Creative Commons BY 3.0 Unported license
© Lucas Wagner

This presentation gives an overview of the qualification of a test case generation tool that utilized model checking to generate tests. The tool is used to satisfy verification objectives, so it was qualified in accordance with DO-330 Tool Qualification Level 5 (TQL-5).

The presentation covers the rationale used for classifying the tool as a TQL-5 tool, the applicable DO-330 objectives for a TQL-5 tool, and examples of how the major objectives were satisfied, including examples of test cases used in the qualification package developed for the test generation tool.

The purpose of this presentation was to give a concrete example and demonstrate that qualification of a tool is not overly complicated, but rather a straightforward, manageable process.

4 Discussion Groups

In addition to individual presentations, the seminar included four discussion groups organized around specific questions that arose during these presentations.

4.1 Why qualify a formal methods tool?

DO-178 (certification standard for software in civil aviation) states that qualification of a tool is needed when certification processes are eliminated, reduced, or automated by the use of a software tool without its output being verified.

For formal methods tools, two questions arise:

- Why use formal methods tools?
- Is qualification necessary?

One difficulty with DO-178 is that structural coverage testing is connected to many different certification objectives. Only some of these objectives can be mitigated using formal methods tools. A careful look at objectives is necessary to determine the economic benefit of using formal methods tools. In some cases, the business case may be derived from a new capability enabled by the use of a formal methods tool. For example:

- The ability to optimize code by using the CompCert compiler (see presentation by Xavier Leroy)
- The ability to increase processor utilization by performing worst case execution time (WCET) analysis with AiT
- The ability to host software at multiple criticality levels on same processor using a verified microkernel such as seL4

Formal methods qualification may, therefore, be a means to justify using the new capability.

Sometimes it is also possible to realize value without qualifying the tool. The use of a formal methods tool to detect and remove errors earlier in the development process is an example. Therefore, the benefit to be derived from a formal methods tool and how it is used in the development process should be carefully evaluated before assuming that qualification is needed.

4.2 How to qualify a formal methods tool?

In this group, we discussed qualification considerations for formal methods tools in the civil aviation context.

DO-333, the formal methods supplement for DO-178C, makes a distinction between a formal method and the tool which implements the method. Additional objectives for formal methods are defined in DO-333 (appearing in tables A-3 through A-5). These objectives apply to the underlying method, and are in addition to any tool qualification activities that may be required. For each formal method used, the following activities should be done:

- Verification that the method has precise unambiguous, mathematically defined syntax and semantic
- Justification of the soundness of the analysis method
- Description and justification of any assumptions that are made in the analysis performed

Concerning tool qualification, there is nothing specific for formal methods tools required by the tool qualification document, DO-330. For verification tools (called TQL-5 tools), the main activities have to do with definition and verification of Tool Operational Requirements. These describe operation of the tool from a user perspective and demonstrate that the tool can satisfy the certification objectives for which it is being used. Some verification must be done showing that the tool does what the requirements say it should do (for example by the use of adequate test cases).

4.3 Compiler qualification strategies

Some formal methods are more difficult to classify in terms of how they fit in to a certification process and what kind of qualification is needed. A good example is the CompCert tool [1]. CompCert is a formally verified C compiler and thus could be seen as a development tool. However, DO-178 is designed to *not* require that the compiler be trusted. Instead, it assumes that executable object code will be verified by means of test (for compliance and robustness with respect to the requirements and to demonstrate structural coverage). The question is thus what is the certification objective that is automated by CompCert?

A possible answer is property preservation between source code and object code. In that case, CompCert could be considered as a verification tool automating this objective, and thus it would be qualified as a TQL-5 tool (according to DO-330). It would, however, be necessary to separate the code production part from the proof part inside the CompCert tool, which is not easy given the nature of the technique used (Coq).

Of course, CompCert could also be qualified as a development tool (TQL-1). In that case, since its assurance story is based on a formal proof, DO-333 (the formal methods supplement to DO-178C) could be applied for the qualification objectives concerning the tool development process. This combination of using formal methods to qualify a formal methods development tool has not been previously considered. In that case, the issue is to justify qualification of CompCert as a development tool on an economic point of view. Since a TQL-1 qualification is costly, it is necessary to determine what can we put in the balance to motivate the use of CompCert in place of a traditional compiler.

References

- 1 Leroy, X. (2009). Formal verification of a realistic compiler. In *Communications of the ACM*, volume 52, number 7, pages 107–115.

4.4 Comparison of qualification in different domains

In this discussion group we discussed the similarities and difference among qualification standards in different domains. The standards considered were:

- DO-178C Software Considerations in Airborne Systems and Equipment Certification and DO-330 Software Tool Qualification Considerations
- IEC 61508 Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 3: Software requirements
- ISO 26262 Road vehicles – Functional safety – Part 8: Supporting processes

The comparison concerned the following questions:

- When is tool qualification required?
- What levels of qualification are defined and what is the purpose of each?
- What activities are required to achieve qualification?

Participants

- June Andronick
UNSW – Sydney, AU
- Rob Arthan
Lemma 1 Ltd. – Twyford, GB
- Jasmin Christian Blanchette
INRIA Lorraine – Nancy, FR
- Sandrine Blazy
IRISA – Rennes, FR
- Matteo Bordin
AdaCore – Paris, FR
- Darren Cofer
Rockwell Collins –
Minneapolis, US
- David Cok
GammaTech Inc. – Ithaca, US
- Rémi Delmas
ONERA – Toulouse, FR
- Michael Dierkes
Rockwell Collins France –
Toulouse, FR
- Eric Engstrom
SIFT – Minneapolis, US
- Gerwin Klein
NICTA – Sydney, AU
- Ramana Kumar
University of Cambridge, GB
- Mark Lawford
McMaster Univ. – Hamilton, CA
- Xavier Leroy
INRIA – Le Chesnay, FR
- Stefan Leue
Universität Konstanz, DE
- Alain Mebsout
Univ. of Iowa – Iowa City, US
- Stephan Merz
INRIA Nancy –
Villers-lès-Nancy, FR
- Cesar A. Munoz
NASA Langley ASDC –
Hampton, US
- Magnus Myreen
University of Cambridge, GB
- Scott Owens
University of Kent, GB
- Marc Pantel
University of Toulouse, FR
- Markus Pister
AbsInt – Saarbrücken, DE
- Werner Schütz
Thales – Wien, AT
- Konrad Slind
Rockwell Collins –
Minneapolis, US
- Nick Tudor
D-RisQ Limited – Malvern, GB
- Lucas Wagner
Rockwell Collins –
Cedar Rapids, US
- Michael W. Whalen
University of Minnesota –
Minneapolis, US
- Virginie Wiels
ONERA – Toulouse, FR

