DAGSTUHL
REPORTS

Volume 4, Issue 9, September 2014

Dagstuhl Reports, Vol. 4, Issue 9 ISSN 2192-5283

ISSN 2192-5283

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

Publication date
January, 2015

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

Digital Object Identifier: 10.4230/DagRep.4.9.i

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

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

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

Editorial Office
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

Report from Dagstuhl Seminar 14371

Adjoint Methods in Computational Science, Engineering, and Finance

Edited by

Nicolas R. Gauger[1], Michael Giles[2], Max Gunzburger[3], and Uwe Naumann[4]

1 **TU Kaiserslautern, DE,** nicolas.gauger@scicomp.uni-kl.de
2 **University of Oxford, GB,** mike.giles@maths.ox.ac.uk
3 **Florida State University, US,** mgunzburger@fsu.edu
4 **RWTH Aachen University, DE,** naumann@stce.rwth-aachen.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 14371 "Adjoint Methods in Computational Science, Engineering, and Finance".

The development of adjoint numerical methods yields a large number of theoretical, algorithmic, and practical (implementation) challenges most of them to be addressed by state of the art Computer Science and Applied Mathematics methodology including parallel high-performance computing, domain-specific program analysis and compiler construction, combinatorial scientific computing, numerical linear algebra / analysis, and functional analysis. One aim of this seminar was to tackle these challenges by setting the stage for accelerated development and deployment of such methods based on in-depth discussions between computer scientists, mathematicians, and practitioners from various (potential) application areas. The number of relevant issues is vast, thus asking for a series of meetings of this type to be initiated by this seminar. It focused on fundamental theoretical issues arising in the context of "continuous vs. discrete adjoints." The relevant context was provided by presentations of various (potential) applications of adjoint methods in CSEF.

Seminar September 8–12, 2014 – http://www.dagstuhl.de/14371
1998 ACM Subject Classification G.1.0 Numerical Algorithms, G.1.4 Automatic Differentiation, G.1.6 Optimization, G.1.7 Ordinary Differential Equations, G.1.8 Partial Differential Equations, G.1.10 Applications, G.4 Mathematical Software
Keywords and phrases continuous adjoints, discrete adjoints, high-performance scientific computing,algorithmic differentiation
Digital Object Identifier 10.4230/DagRep.4.9.1

1 Executive Summary

Uwe Naumann

The human desire for meaningful numerical simulation of physical, chemical, biological, economical, financial (etc.) phenomena in CSEF has been increasing with the growing performance of the continuously improving computer systems. As a result of this development we are (and will always be) faced with a large (and growing) number of highly complex numerical simulation codes that run at the limit of the available HPC resources. These codes often result from the discretization of systems of PDE. Their run time correlates with the spatial and temporal resolution which often needs to be very high in order to capture the real behavior of the underlying system. There is no doubt that the available hardware will always be used to the extreme. Improvements in the run time of the simulations need to be sought through research in numerical algorithms and their efficient implementation on HPC architectures.

Problem sizes are often in the billions of unknowns; and with emerging large-scale computing systems, this size is expected to increase by a factor of thousand over the next five years. Moreover, simulations are increasingly used in design optimization and parameter identification which is even more complex and requires the highest possible computational performance and fundamental enabling algorithmic technology. Derivatives of certain objectives of these numerical models with respect to a potentially very large number of model parameters are crucial for the highly desirable transition from pure simulation to optimization. Approximation of these derivatives via finite difference quotients often lacks the required accuracy. More importantly, it may be infeasible for a large parameter space in terms of its computational complexity. Adjoint numerical programs have until recently been written by hand to overcome this problem. Such programs compute (large) gradients with machine accuracy at a small constant multiple of the computational complexity of the underlying primal simulation. Due to the enormous size of most numerical simulation codes the manual procedure may take up to several man years. Moreover manual adjoint codes are error-prone and hard to maintain as the primal simulation evolves. Computer scientists have been developing special software tools based on the principles of algorithmic differentiation (AD) to generate discrete adjoint code automatically. Consequently, this method has gained considerable acceptance within the CSEF community as illustrated by numerous successful case studies presented in the proceedings of so far six international conferences on AD. See http://www.autodiff.org for details.

Illustrative Example: Classical applications of adjoint methods arise in the context of large-scale inverse problems, such as the estimation of unknown or uncertain parameters of implementations of mathematical models for real-world problems as computer programs. Imagine the optimization of the shape of an aircraft with the objective to maximize its lift. The continuous mathematical domain (the surface of the aircraft) is typically discretized through the generation of a mesh with a potentially very large number of points spread over the whole surface. Optimization aims to adapt the position of these points in 3D space such that the objective is met while at the same time satisfying various constraints (e. g. prescribed volume). A naive approach might run a potentially very large number of primal numerical simulations with changing mesh configurations thus being able to identify an optimum within this very limited search space.

Derivative-based approaches use information on the sensitivity of the objective at the given mesh configuration with respect to changes in the positions of all mesh points (the gradient) in order to make a deterministic decision about the next configuration to be considered. The sensitivities can be approximated through local perturbations of the position of each mesh point (finite difference quotients). A single optimization step would thus require a number of primal simulations that is of the order of the number of degrees of freedom (three spatial coordinates for each mesh point) induced by the mesh. This approach is practically infeasible as a single simulation may easily run for several minutes (if not hours) on the latest HPC architectures. The approximation of a single gradient would take months (if not years) for a mesh with only one million points.

Adjoint methods deliver the gradient at the cost of only a few (between 2 and 10) primal simulations. Continuous adjoint methods derive an adjoint version of the primal mathematical model analytically followed by the numerical solution of the resulting adjoint model. While this approach promises low computational cost (approx. 2 primal simulations) it can be mathematically challenging and numerically inconsistent when compared with the primal numerical simulation. To the best of our knowledge, the automation of the derivation of continuous adjoint models is still outstanding.

Discrete adjoint methods rely on the algorithmic differentiation of the primal numerical model, thus overcoming the potential numerical inconsistencies induced by the continuous adjoint. Depending on the mode of implementation of AD, the level of maturity of the AD tool, and the expertise of the user of the tool the computational cost can range between 2 and 20 primal simulations, sometimes even more. Still this cost is independent of the number of mesh points (referring to the above example). Solutions to problems arising in adjoint methods require expertise in both theoretical and applied Computer Science as well as in Numerical Analysis. Robust methods for the data flow reversal within adjoint code are built on special graph partitioning and coloring algorithms. Their implementation on modern HPC architectures (e. g. using MPI and/or OpenMP) has impact on the simulation software design and the data management. The use of accelerators has been considered only recently with many open as of yet unsolved problems. Static and dynamic program analysis and compiler construction techniques have been developed to facilitate the semi-automatic generation of discrete adjoint code. The exploration of a potential extension of these techniques to continuous adjoint code was one of the subjects of this seminar. Other conceptual problems discussed included functional analytic aspects of adjoint methods and their impact on practical implementation, combinatorial problems in adjoint code generation and their computational complexities, and simulation software engineering guidelines in the light of adjoint methods.

Adjoint methods borrow from a variety of subfields of Computer Science and Applied Mathematics including high performance and combinatorial scientific computing, program analysis and compiler construction, functional analysis, numerical analysis and linear algebra, and with relevance to a wide range of potential areas of application. As such, the topic lends itself to a series of seminars taking more detailed looks into the respective subjects. With this seminar we intent to initiate a sequence of related events alternating in between the Leibniz Center for Informatics at Schloss Dagstuhl and the Mathematisches Forschungsinstitut Oberwolfach, thus, emphasizing the obvious synergies between Computer Science and Mathematics in the given context.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 OO-Lint for Operator Overloading in C++

Christian Bischof (TU Darmstadt, DE)

Automatic Differentiation Tools in C++ often employ operator overloading (OO), and in particular almost all reverse mode tools employ this technology. This seemingly simple approach just requires a type change in the numeric code, redeclaring the floating point type. However, in practice, it is not so simple, as the operator overloading may lead to conflicts with the C++ standard, such as, for example, multiple user defined conversions, unions with complex types, or implicit conversions in conditions, and, as a result, to cryptic error messages.

To alleviate this problem, we developed a tool based on LLVM/Clang, which recognizes problematic coding constructs in the C++ code to be subjected to operator overloading. This tool thus provides guidance to a potential user of OO-based semantic enhancements of an existing code, in a fashion that is much more targeted and understandable than the usual error messages produced by the compiler. In addition, we are working on automating the changes necessary to make the code applicable to OO, thus further easing the transition for potential users of OO-based AD tools.

3.2 An extension of the projected gradient method with application in multi-material structural topology optimization

Luise Blank (Universität Regensburg, DE)

Joint work of Blank, Luise; Rupprecht, Christoph

First we introduce the phase field approach for the optimal distribution of several elastic isotropic homogeneous materials. This leads to a multi-material structural topology optimization problem with linear elasticity equations, mass constraints and pointwise inequalities as restrictions. The reduced problem formulation results into a nonlinear optimization problem over a convex and closed set. While the reduced cost functional is Fréchet-differentiable in $H^1 \cap L^\infty$ it is not differentiable in a Hilbert-space. Hence the classical theory for projected gradient methods cannot be applied. Therefore, we extend the projected gradient method to Banach spaces. The gradient is not required but only directional derivatives. Furthermore, variable scaling and varying the metric is introduced. The last allows the use of second order information in the method. We prove global convergence of the method. This method is applied to the presented optimization problem. Here it turns out that the scaling of the derivative with respect to the interface thickness is important to obtain a drastic speed up of the method. With computational experiments we demonstrate the independence of the discretization mesh size and of the interface thickness in the number of iterations as well as its efficency in time. Moreover we present results for compliance mechanism and drag minimization in Stokes flow.

3.3 Algorithmic Differentiation for Geometry Processing

David Bommes (INRIA Sophia Antipolis – Méditerranée, FR)

Joint work of Bommes, David; Lotz, Johannes; Naumann, Uwe

Recent geometry processing approaches are often related to optimization of complicated nonlinear functionals and constraints. The resulting problems are usually optimized with interior point methods that require second order derivative information. The goal of this project is to develop a framework for rapid prototyping of such applications with the help of algorithmic differentiation, where based on a user-provided functional evaluation, all derivatives should be generated automatically. To obtain a performance that is comparable to the manual approach, it is crucial to exploit the sparsity structure that geometry processing approaches provide due to partial separability [1] of the corresponding functionals. By providing an intuitive interface where the partial separability becomes obvious, we develop a fast an easy to use framework for geometry processing with AD, which integrates C++ overloading techniques like ADOL-C [2] and dco/c++ [3].

References

1 Bischof, C. H., Bouaricha, A., Khademi, P. M., Moré, J. J. (1997). Computing gradients in large-scale optimization using automatic differentiation. INFORMS Journal on Computing, 9(2), 185–194.

2 Walther, A., Griewank, A. (2012). Getting started with ADOL-C. Combinatorial Scientific Computing, 181–202.

3 Naumann, U. (2011). The art of differentiating computer programs: an introduction to algorithmic differentiation (Vol. 24). SIAM.

3.4 An interface for conveying high-level user information to automatic differentiation transformations: A working group proposal

Martin Bücker (Universität Jena, DE)

Program transformations that augment a given computer code with statements for the computations of derivatives are commonly referred to as automatic or algorithmic differentiation (AD). Software tools implementing the AD technology are available for various programming languages; see the community web site www.autodiff.org. Today, there are robust AD tools which are capable of correctly transforming large programs with minimal human intervention. However, sometimes, AD is not “automatic.” In fact, a combined approach that applies AD in a black-box fashion to large parts of the code and that also involves a moderate amount of human intervention for certain parts of the code is often adequate. This Dagstuhl seminar is a perfect opportunity to discuss the following questions: Where is human intervention absolutely necessary? What can be handled mechanically by an AD tool? To what extent can the level of abstraction be raised by user-specified directives?

3.5 The Glorious Future of Automatic Differentiation: A Retrospective View

Bruce Christianson (University of Hertfordshire, GB)

The adjoint mode of AD is both fast and accurate. We have had considerable success with selling the advantage of speed to a variety of application communities. To a large extent this is because AD tools have evolved so as to cope well with legacy code. There is still room for improvement, but we should not become fixated on embracing the legacy agenda.

With regard to the second advantage, we have made little headway in selling accuracy as a step-change benefit. This failure is largely due to the continued use of legacy approaches to modeling, and to optimization.

At present, a smooth continuous model is typically discretized, not always consistently, before being implemented (using loops with an epsilon-based stopping criterion) into a program that is not continuous, let alone smooth. Finite difference approximations are used to smooth these discontinuities over the scale of the anticipated step. The resulting values are used by optimization algorithms (such as Quasi-Newton), to build internally a new smooth local model, that is inconsistent with the original model. Finally, this new model is solved exactly.

It is clear that simply differentiating the discontinuous program accurately does not much help overall performance or convergence: users want secants not tangents; the primal problem is seldom converged accurately until near the optimum; the raw models exhibit chaotic behavior; and Newton is not stable.

Louis Rall often pointed out that Automatic Differentiation is not really a local operation. What mileage can we gain by re-factoring AD to obtain accurate, and more importantly, consistent solutions to systematic smooth perturbations of the original model, for example?

As a side-effect, the adjoint mode produces large numbers of potentially useful by-products, such as Lagrange multipliers, for free. However, when solving equations, legacy modeling software often does not explicitly identify, or even calculate, the corresponding equation residuals.

We, the AD community, need urgently to clarify our thinking, and prepare our agenda for changing the next generation of modeling and optimization tools.

3.6 Adjoint-Based Research and Applications at NASA Langley Research Center

Boris Diskin (National Institute of Aerospace – Hampton, US)

Joint work of Diskin, Boris; Nielsen, Eric
Main reference E. Nielsen, B. Diskin, "Discrete Adjoint-Based Design for Unsteady Turbulent Flows on Dynamic Overset Unstructured Grids," AIAA Journal, 51(6):1355–1373, June 2013; available also from NASA repository.
URL http://dx.doi.org/10.2514/1.J051859
URL http://fun3d.larc.nasa.gov/papers/overset_adjoint_aiaaj.pdf

An overview of the use of adjoint methods at NASA Langley Research Center has been presented. Among major areas of large-scale adjoint-based applications are shape optimization, grid adaptation, and multidisciplinary optimization. In this talk, examples of unsteady adjoint-based aerodynamic shape optimization have been presented such as active flow control for a high-lift configuration, a helicopter in forward flight, a fighter jet with simulated aeroelastic effects, and a biologically-inspired flapping wing configuration. Several examples of adjoint-based mesh adaptation for various applications have included high-lift and nozzle plume configurations, a sonic boom application, an example of shock-boundary layer interaction, and several other aerospace applications. Multidisciplinary optimization capabilities have been demonstrated for sonic boom mitigation. High-performance computing aspects have also been discussed, such as scaling performance in the presence of frequent I/O for unsteady adjoint simulations.

3.7 Automated adjoint finite element simulations within FEniCS

Patrick Farrell (University of Oxford, GB)

Joint work of Farrell, Patrick; Funke, Simon; Ham, David; Rognes, Marie; Birkisson, Ásgeir
Main reference P. E. Farrell, D. A. Ham, S. W. Funke, M. E. Rognes, "Automated derivation of the adjoint of high-level transient finite element programs," SIAM Journal on Scientific Computing, 35(4):C369–C393, July 2013.
URL http://dx.doi.org/10.1137/120873558

In this work, we develop and advocate a high-level approach to automated adjoint derivation for finite element simulators. By "high-level", we mean that instead of breaking down a program into a sequence of elementary machine instructions, we treat the program in terms of much higher-level mathematical constructs: in the finite element case, the solution of variational problems. By retaining as much mathematical structure as possible, this approach offers several advantages: the derived tangent linear and adjoint solutions work naturally in parallel; the adjoint solver can automatically use optimal checkpointing schemes with minimal user intervention; and the tangent linear and adjoint models typically exhibit optimal efficiency.

This high-level perspective permits optimizations and modifications that would be impractical in low-level code. For example, when computing Hessian actions, the equations to be solved for each action (the tangent linear and second-order adjoint equations) share the same matrices to be solved, up to transposition: a high-level AD tool can cache the factorizations of these matrices and re-use them to dramatically speed up the per-action cost, but such an optimization would be extremely difficult to implement in a low-level AD tool.

Several examples were presented, including nonlinear diffusion on a manifold; Hessian eigendecomposition for the test case of Deckelnick and Hinze; and mesh independence in solving the mother problem of PDE-constrained optimization.

In general, to attain optimal performance with algorithmic differentiation, the tool needs to exploit all of the mathematical structure available in a given problem domain. This work is specific to finite elements, but it would be entirely possible to apply the experience gained to finite volume discretizations, or problem domains outside of PDEs.

In addition, I presented very recent results on deflation techniques for computing distinct solutions of nonlinear systems and distinct local minima of nonconvex optimization problems.

3.8 Physical Interpretations of Discrete and Continuous Adjoint Boundary Conditions

Christian Frey (DLR – Köln, DE)

Joint work of Frey, Christian; Engels-Putzka, Kügeler, Edmund
Main reference C. Frey, A. Engels-Putzka, E. Kügeler, "Adjoint boundary conditions for turbomachinery flows," in Proc. of the 2012 Europ. Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS'12), pp. 9360–9379, 2012.
URL http://elib.dlr.de/81285/

In this talk I have discussed the treatment of boundary conditions in the context of a discrete adjoint industrial turbomachinery RANS solver. Special emphasis is put on the non-reflecting boundary conditions and the blade row coupling by mixing planes. These techniques are widely used for the accurate approximation of time-averaged flows in turbomachinery by steady simulations. In contrast to inviscid wall boundary conditions these boundary conditions are applied at non-characteristic boundaries, i. e., the flux Jacobian through the boundary is non-singular, unless the normal flow vanishes at some point of the blade row entry or exit. On the other hand, these boundary conditions are more complicated in that they are non-local. They involve, for instance, Fourier transformations, and special averaging techniques.

We outline a general methodology to adjoin discretely numerical boundary conditions and apply the techniques to the boundary conditions of an internal flow solver. This leads to adjoint boundary update operators which are applied after each multiplication with the adjoint residual Jacobian. This methodology carries over to the communication using domain decomposition and ghost cells.

A further difficulty that one has to deal with is the fact that the boundary conditions in the non-linear solver are implemented as fixed-point iterations, whereas the adjoint system is solved by the GMRES method. This means that the application of AD seems to be rather difficult and may require that one switches to a discrete adjoint iterative solver.

The second part of this talk is dedicated to the physical interpretation of the discrete adjoint boundary update as discretizations of their continuous adjoint counterpart. For this purpose we determine the continuous adjoint boundary conditions for the above-mentioned turbomachinery boundary conditions. Finally we give a physical interpretation of the adjoint non-reflecting condition and the adjoint mixing plane coupling condition. The former can be viewed as a non-reflecting condition for the adjoint Euler equations. The adjoint blade row coupling condition is satisfied if certain circumferential averages of the adjoint fields on both sides of the interfaces agree up to a factor given by the blade count ratio.

3.9 Efficient Adjoint-based Techniques for Optimal Active Flow Control

Nicolas R. Gauger (TU Kaiserslautern, DE)

Joint work of Gauger, Nicolas R.; Wang, Qiqi; Kramer, Felix; Thiele, Frank; Günther, Stefanie; Sagebaum, Max

For efficient optimal active control of unsteady flows, the use of adjoint approaches is a first essential ingredient. We compare continuous and discrete adjoint approaches in terms of accuracy, efficiency and robustness. For the generation of discrete adjoint solvers, we discuss the use of Automatic Differentiation (AD) and its combination with checkpointing techniques. Furthermore, we discuss so-called one-shot methods. Here, one achieves simultaneously convergence of the primal state equation, the adjoint state equation as well as the design equation. The direction and size of the one-shot optimization steps are determined by a carefully selected design space preconditioner. The one-shot method has proven to be very efficient in optimization with steady partial differential equations (PDEs). Applications of the one-shot method in the field of aerodynamic shape optimization with steady Navier-Stokes equations have shown, that the computational cost for an optimization, measured in runtime as well as iteration counts, is only 2 to 8 times the cost of a single simulation of the governing PDE. We present a framework for applying the one-shot approach also to optimal control problems with unsteady Navier-Stokes equations. Straight forward applications of the one-shot method to unsteady problems have shown, that its efficiency depends on the resolution of the physical time domain. In order to dissolve this dependency, we consider unsteady model problems and investigate an adaptive time scaling approach.

3.10 Integrating and Adjoining ODEs with Lipschitzian RHS

Andreas Griewank (HU Berlin, DE)

Joint work of Griewank, Andreas; Lenser, Ben; Hasenfelder, Richard; Streubel,Tom; Boeck, Paul; Gompil, Battur
Main reference P. Boeck, B. Gompil, A. Griewank, R. Hasenfelder, N. Strogies, "Experiments with generalized midpoint and trapezoidal rules on two nonsmooth ODE's," Mong. Math. J., 17:39–49, 2013.
URL http://iom.num.edu.mn/journal/2013/3.pdf

We consider initial value problems in ODEs where the right hand side has truly state-dependent kinks or jumps and users may be hard pressed to provide suitable switching functions for the customary event handling approaches. Instead kinks and jumps can be detected and handled automatically, which is possible by an extension of algorithmic differentiation that provides piecewise linear approximations with second order error to piecewise smooth and Lipschitz continuous right hand sides [3]. Without continuity of the underlying piecewise smooth function the resulting piecewise linear approximation will also be discontinuous and the approximation error is no longer uniform but heavily direction dependent. Nevertheless we expect to extend our approach later to ODEs where the RHS has jumps but the exact solution trajectories satisfy a certain transversality condition. In the Lipschitzian case we show how the piecewise linearizations of the RHS is generated by Algorithmic Piecewise Differentiation abs-normal form and how it can be used to generalize the midpoint and the trapezoidal rule such that local third order consistency and uniform global convergence order two is recovered [1]. The inherent smoothness of the approximation facilitates the gain of one or two extra orders by Richardson/Romberg extrapolation. The two implicit discretizations of the ODEs produce non-smooth systems of algebraic equations which have been solved by the methods discussed in [4]. The corresponding adjoint trajectory is defined by a differential inclusion and thus not unique if there are valley tracing modes as defined by P. Barton and K. Khan [2]. Nevertheless, usually one obtains a generalized gradient that can be used for data assimilation and more general optimal control. We report preliminary results on a shallow water equation in 1D where non smoothness arises through slope limiters or no smooth norms in the error functional.

References

1 P. Boeck, B. Gompil, A. Griewank, R. Hasenfelder, and N. Strogies*Experiments with generalized midpoint and trapezoidal rules on two nonsmooth ODE's*, Mong. Math. J., Vol. 17, pp. 39–49. (2013)

2 Kamil A. Khan, Paul I. Barton *Switching behavior of solutions of ordinary differential equations with nonsmooth right-hand sides* (2014)

3 Andreas Griewank. *On Stable Piecewise Linearization and Generalized Differentiation.* Optimization Methods and Software, 28(6):1139–1178 (2013)

4 A. Griewank, J.-U. Bernt, M. Radons, and T. Streubel. *Solving piecewise linear equations in abs-normal form.* Optimization-Online 2013/12/4184 (2013).

3.11 Discrete versus Continuous Adjoints of Differential-Algebraic Equation Systems: Similarities and Differences

Ralf Hannemann-Tamas (Univ. of Science & Technology – Trondheim, NO)

Main reference Ralf Hannemann-Tamas, "Adjoint Sensitivity Analysis for Optimal Control of Non-Smooth Differential-Algebraic Equations," Dissertation, Shaker Verlag, 2013.
URL http://www.shaker.de/de/content/catalogue/index.asp?lang=de&ISBN=978-3-8440-1635-2

This work is based on the PhD thesis [1]. In contrast to multi-step methods [2], one step-methods applied to semi-explicit DAE systems of index 1, result in differential discrete adjoints which are consistent with the differential continuous adjoints, while the discrete adjoints for algebraic variables tend to zero along their trajectory.

We illustrate this fact by a small example. Let R denote the field of the real numbers and let $\psi : R^{n_x} \to R$ be a smooth function. We aim to compute the gradient of the scalar functional J, where $J(x, y, p) := \psi(x(t_1))$, with respect to the parameter vector p, where $x(t_1)$ is characterized by the parametric initial value problem

$$\begin{aligned} \dot{x} &= f(x, y), \\ 0 &= g(x, y), \\ x(t_0) &= p, \\ y(t_0) &= y_0. \end{aligned}$$

Here, $x(t) \in R^{n_x}$, $y(t) \in R^{n_y}$, $p \in R^{n_p}$ denote the differential variables, algebraic variables and the parameters, respectively. The initial and final times are t_0 and t_1, respectively and the mappings $f : R^{n_x} \times R^{n_y} \to R^{n_x}$ and $g : R^{n_x} \times R^{n_y} \to R^{n_y}$ are assumed to be sufficiently smooth. Further, the initial value problem is assumed to have a unique solution.

For the moment we assume that initial values y_0 are consistent, i. e., they are a solution of the algebraic equation

$$g(p, y_0) = 0.$$

Let x_1, y_1 approximate the differential and algebraic variables at time t_1 as the solution of the one-step method (in Henrici's notation)

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} p \\ y_0 \end{pmatrix} + \Phi(t_0, x_0, y_0, h), \quad \text{with} \quad h = t_1 - t_0.$$

Then, the discrete adjoints λ_0^x, λ_0^y at time t_0 can be interpreted as derivatives of the objective with respect to the initial values

$$\lambda_0^x = \frac{\partial\psi(x_1)}{\partial p}, \quad \lambda_0^y = \frac{\partial\psi(x_1)}{\partial y_0}. \tag{1}$$

Especially, the adjoints λ_0^y, associated with the algebraic variables, describe the sensitivity of the objective function with respect to perturbations δy_0 of the then inconsistent initial values $\lambda_0^y = y_0 + \delta y_0$. However, a good numerical method evens out small inconsistencies of algebraic initial values. Hence, the values of the algebraic adjoints λ_0^y tend to zero, i. e. $\lambda_0^y \approx 0$. In particular, at $t = t_0$, the discrete adjoints λ_0^y deviate from continuous algebraic adjoints $\lambda^y(t_0)$, since they are usually different from zero (e. g. see [3]). In contrast, the differential discrete adjoints λ_0^x usually converge with high order against the differential continuous adjoints $\lambda^x(t_0)$, since these satisfy (e. g. see [3])

$$\lambda^x(t_0) = \frac{\partial\psi(x(t_1))}{\partial p},$$

which is analogous to the first identity in eq. (1).

To summarize, the similarity is that differential discrete adjoints converge (for $h \to 0$) to their continuous counterparts. The difference is, that the latter convergence result does not hold for algebraic discrete adjoints.

References

1 R. Hannemann-Tamás. *Adjoint Sensitivity Analysis for Optimal Control of Non-Smooth Differential-Algebraic Equations.* Dissertation, Shaker Verlag, 2013.

2 A. Sandu. *Reverse automatic differentiation of linear multistep methods.* In T. J. Barth, M. Griebel, D. E. Keyes, R. M. Nieminen, D. Roose, T. Schlick, C. H. Bischof, H. M. Bücker, P. Hovland, U. Naumann, and J. Utke (Eds.), *Advances in Automatic Differentiation*, Volume 64 of Lecture Notes in Computational Science and Engineering, pp. 1–12. Springer, 2008.

3 Y. Cao, S. Li, L. Petzold, and R. Serban. *Adjoint sensitivity analysis for differential-algebraic equations: The adjoint DAE system and its numerical solution.* SIAM J. Sci. Comput. *24*(3), pp. 1076–1089, 2003.

3.12 Bridges between OO-based and ST-based adjoint AD

Laurent Hascoet (INRIA Sophia Antipolis – Méditerranée, FR)

Among AD tools, there is a clear opposition between those based on Operator Overloading (OO) and those based on Source Transformation (ST). Competition between the two classes of tools has brought huge improvements to both but obviously no class will ever show definitive superiority. We claim that this competition, fruitful as it was, must give way to collaboration between OO-based and ST-based AD models.

We believe that the difference between OO-based and ST-based adjoint AD is not as deep as one may think at first sight. Indeed we think that the vocabulary and techniques of Partial Evaluation can help us exhibit a close resemblance, not only at some conceptual level, but also leading to fruitful exchange of techniques between OO-based and ST-based.

In the framework of Partial Evaluation, we can view the tape, built by the tape-recording phase of OO-based adjoint differentiation, as made of two parts, one part being static i. e. depending only on the program to differentiate, the other being dynamic i. e. depending also on the particular input. Obviously the static part could be extracted once and for all from the program to differentiate. By Partial Evaluation of the Tape Interpreter with respect to the static part of the tape, one obtains a Specialized Tape Interpreter that, when given the dynamic part, will evaluate the adjoint derivatives more efficiently than the initial Tape Interpreter. Also, the dynamic part of the Tape is smaller than the full tape, and can actually be much smaller. We claim that the computation of the dynamic part of the tape on one hand and the execution of the Specialized Tape Interpreter on the other hand, correspond exactly to the two phases of a ST-based adjoint namely, the forward sweep on one hand and the backward sweep on the other hand.

Practically, we think we tool developers should explore more ways to make OO-based and ST-based AD tools collaborate. One key architectural choice being "who is in the driver's seat". For instance black-box mechanisms in OO-based environments allow them to call ST-based adjoints for computational kernels where the language constructs pose no difficulty to static source analysis and transformation. Only this is still a tedious process. Conversely, there must be ways for OO-based AD to take better advantage of the static data-flow analysis (activity, liveness, TBR, . . .) that ST-based AD computes and uses routinely. In particular the activity analysis of an ST-based tool can be used to automate the type transformation stage that an OO-based AD used must perform by hand.

3.13 Application of derivative code in climate modeling

Patrick Heimbach (MIT – Cambridge, US)

Main reference C. Wunsch, P. Heimbach, "Dynamically and kinematically consistent global ocean circulation and ice state estimates," chapter 21 in G. Siedler, J. Church, J. Gould, S. Griffies, eds., "Ocean Circulation and Climate: A 21st Century Perspective," pp. 553–579, Elsevier, 2013.
Main reference http://store.elsevier.com/product.jsp?isbn=9780123918536
URL http://ecco-group.org

Optimal state and parameter estimation, accompanied by rigorous uncertainty quantification, is increasingly being recognized as a powerful tool in climate modeling. The need arises in order to deal with the problem of sparse observations to optimally constrain model simulations and infer dynamically consistent time-evolving state estimates (Heimbach and Wunsch 2012), to provide quantitative estimates of the extent to which existing observations constrain uncertain parameters or the optimal design of future observing networks (Heimbach et al. 2010), or to infer optimal initial conditions that are best suited for predictions or projections (Zanna et al. 2012). The primary example given is that of estimating the global ocean (and sea ice) circulation over the last few decades as undertaken by the "Estimating the Circulation and Climate of the Ocean" (ecco-group.org) consortium (Stammer et al. 2002; Wunsch and Heimbach 2007, 2013, 2014).

A number of regional efforts targeting higher spatial resolutions and shorter time scales are also being pursued to synthesize the available data with the known dynamics, e. g., in the North Atlantic (Ayoub 2006; Gebbie et al. 2006), the Southern Ocean (Mazloff et al. 2010), the tropical Pacific (Hoteit et al. 2010), or the Labrador Sea and Baffin Bay (Fenty and Heimbach 2013).

Similar efforts by a number of groups are now targeting the polar ice sheets for the purpose of developing predictive capabilities and uncertainty estimates of ice sheet mass loss in the coming centuries (Heimbach and Bugnion 2009; Goldberg and Heimbach 2013; Larour et al. 2014; Perego et al. 2014).

Increasingly, second derivative, i. e. Hessian information is being explored to provide a posteriori uncertainty estimates on addition to the maximum a posteriori probability estimate, and to propagate these uncertainties forward onto target quantities of interest, e. g. climate indices (Kalmikov and Heimbach 2014; Petra et al. 2014).

In order to improve accessibility to algorithmic differentiation (AD) tools that support flexible derivative code development, the ECCO group has been involved in the development, at Argonne National Lab, of the open source AD tool OpenAD (Naumann et al. 2006; Utke et al. 2008, 2009). Today, a number of configurations of the MIT general circulation model (mitgcm.org) are available for AD-enhanced simulations.

References

1 Ayoub, N. (2006). Estimation of boundary values in a North Atlantic circulation model using an adjoint method. Ocean Modeling, 12(3-4), 319–347. doi:10.1016/j.ocemod.2005.06.003

2 Fenty, I.G. and P. Heimbach (2013). Coupled Sea Ice-Ocean State Estimation in the Labrador Sea and Baffin Bay. J. Phys. Oceanogr., 43(6), 884–904, doi:10.1175/JPO-D-12-065.1.

3 Gebbie, G., P. Heimbach and C. Wunsch, 2006: Strategies for nested and eddy-resolving state estimation. J. Geophys. Res., 111, C10073, doi:10.1029/2005JC003094.

4 Goldberg, D.N. and P. Heimbach (2013). Parameter and state estimation with a time-dependent adjoint marine ice sheet model. The Cryosphere, 7, 1659–1678, doi:10.5194/tc-7-1659-2013.

5 Heimbach, P. and C. Wunsch (2012). Decadal ocean (and ice) state estimation for climate research: What are the needs? Oberwolfach Reports, 9(4), 3451–3454, doi:10.4171/OWR/2012/58

6 Heimbach, P. and V. Bugnion (2009). Greenland ice sheet volume sensitivity to basal, surface, and initial conditions, derived from an adjoint model. Annals of Glaciology, 50(52), 67–80, doi:10.3189/172756409789624256

7 Heimbach, P., G. Forget, R. Ponte, and C. Wunsch, et al. (2010). Observational Requirements for global-scale ocean climate analysis: Lessons from ocean state estimation. Community White Paper. In: Hall, J., D. E. Harrison, and D. Stammer (Eds.), 2010: Proc. of OceanObs 09: Sustained Ocean Observations and Information for Society, Venice, Italy, 21–25 September 2009, ESA Publication WPP-306, Vol. 2, doi:10.5270/OceanObs09.cwp.42.

8 Heimbach, P., C. Hill and R. Giering (2002). Automatic Generation of Efficient Adjoint Code for a Parallel Navier-Stokes Solver. in: J.J. Dongarra, P.M.A. Sloot and C.J.K. Tan (Eds.), Lecture Notes in Computer Science (LNCS), Vol. 2330, part II, pp. 1019–1028, Springer-Verlag, doi:10.1007/3-540-46080-2_107.

9 Hoteit, I., B. Cornuelle, and P. Heimbach, 2010: An Eddy-Permitting, Dynamically Consistent Adjoint-Based Assimilation System for the Tropical Pacific: Hindcast Experiments in 2000. J. Geophys. Res., 115, C03001, doi:10.1029/2009JC005437.

10 Kalmikov, A. and P. Heimbach (2014). A Hessian-based method for Uncertainty Quantification in Global Ocean State Estimation. SIAM J. Scientific Computing (Special Section on Planet Earth and Big Data), 36(5), S267–S295, doi:10.1137/130925311.

11 Larour, E., Utke, J., Csatho, B., Schenk, A., Seroussi, H., Morlighem, M., Rignot, E., Schlegel, N., and Khazendar, A. (2014). Inferred basal friction and surface mass balance of North-East Greenland Ice Stream using data assimilation of ICESat-1 surface altimetry and ISSM, The Cryosphere Discuss., 8, 2331–2373, doi:10.5194/tcd-8-2331-2014.

12 Mazloff, M.R., P. Heimbach and C. Wunsch (2010). An Eddy-Permitting Southern Ocean State Estimate. J. Phys. Oceanogr., 40(5), 880–899, doi:10.1175/2009JPO4236.1.

13 Naumann, U., J. Utke, P. Heimbach, C. Hill, D. Ozyurt, C. Wunsch, M. Fagan, N. Thallent, M. Strout (2006). Adjoint code by source transformation with OpenAD/F. In: Proc. of the Europ. Conf. on Computational Fluid Dynamics (ECCOMAS CFD 2006). P. Wesseling, J. Periaux, and E. Onate (Eds.), TU Delft, The Netherlands, 2006.

14 Perego, M., Price, S., and Stadler, G. (2014). Optimal initial conditions for coupling ice sheet models to Earth system models. Journal of Geophysical Research: Earth Surface, 119(9), 1894–1917. doi:10.1002/2014JF003181

15 Petra, N., Martin, J., Stadler, G., and Ghattas, O. (2014). A Computational Framework for Infinite-Dimensional Bayesian Inverse Problems, Part II: Stochastic Newton MCMC with Application to Ice Sheet Flow Inverse Problems. SIAM Journal on Scientific Computing, 36(4), A1525–A1555. doi:10.1137/130934805

16 Stammer, D., C. Wunsch, R. Giering, C. Eckert, P. Heimbach, J. Marotzke, A. Adcroft, C. Hill, J. and J. Marshall (2002). The global ocean circulation during 1992–1997, estimated from ocean observations and a general circulation model. J. Geophys. Res., 107 (C9), 3118, doi:10.1029/2001JC000888

17 Utke, J., L. Harscoet, P. Heimbach, C. Hill, P. Hovland and U. Naumann (2009). Toward adjointable MPI. Proceedings of the 10th IEEE International Workshop on Parallel and Distributed Scientific and Engineering, PDSEC-09, Rome, Italy, pp. 1–8, doi:10.1109/IPDPS.2009.5161165.

18 Utke, J., U. Naumann, M. Fagan, N. Thallent, M. Strout, P. Heimbach, C. Hill and C. Wunsch (2008). OpenAD/F: A modular, open-source tool for automatic differentiation of Fortran codes. ACM Transactions on Mathematical Software (TOMS), 34(4), doi:10.1145/1377596.1377598

19 Wunsch, C. and P. Heimbach (2014). Bidecadal Thermal Changes in the Abyssal Ocean. J. Phys. Oceanogr., 44(8), 2013-2030, doi:10.1175/JPO-D-13- 096.1.

20 Wunsch, C. and P. Heimbach (2013). Dynamically and kinematically consistent global ocean circulation and ice state estimates. In: G. Siedler, J. Church, J. Gould and S. Griffies, eds.: Ocean Circulation and Climate: A 21st Century Perspective. Chapter 21, pp. 553–579, Elsevier, doi:10.1016/B978-0-12- 391851-2.00021-0.

21 Wunsch, C. and P. Heimbach (2007). Practical global ocean state estimation. Physica D, 230(1-2), pp. 197–208, doi:10.1016/j.physd.2006.09.040.

22 Zanna L., P. Heimbach, A. M. Moore and E. Tziperman (2012). Upper-ocean singular vectors of the North Atlantic climate with implications for linear predictability and variability. Quart. J. Roy. Met. Soc., 138(663), 500–513, doi:10.1002/qj.937.

3.14 Adjoints in Solution Methods for PDE Constrained Inverse Problems: Reduced versus all-at-once formulations

Barbara Kaltenbacher (Universität Klagenfurt, AT)

Main reference B. Kaltenbacher, A. Kirchner, B. Vexler, "Goal oriented adaptivity in the IRGNM for parameter identification in PDEs II: all-at-once formulations," Inverse Problems, 30(4):045002, 2014
URL http://dx.doi.org/10.1088/0266-5611/30/4/045002

Inverse problems for partial differential equation such as identification of coefficients, boundary conditions or source terms appear in a wide range of applications. Due to their inherent instability, regularization has to be applied. When computing regularized solutions, adjoints

naturally appear: In minimization based regularization methods like Tikhonov within the first order optimality conditions; for iterative (Newton type or gradient) regularization methods directly in the definition of each step.
The system from which parameters are to be identified typically consist of two parts: The model equation, e. g. a (system of) ordinary or partial differential equation(s), and the observation equation. In the conventional reduced setting, the model equation is eliminated via the parameter-to-state map. Alternatively, one might consider both sets of equations (model and observations) as one large system, to which some regularization method is applied. The choice of the formulation (reduced or all-at-once) can make a large difference computationally, depending on which regularization method is used: Whereas almost the same optimality system arises for the reduced and the all-at-once Tikhonov method, the situation is different for Landweber (LW), i. e., gradient methods: A reduced LW iteration requires solution of the PDE and the (linear) adjoint in each step, whereas in all-at-once LW, only PDE residuals have to be evaluated, but no PDEs need to be solved. In between lie Newton type methods, whose reduced versions again need PDE and adjoint PDE solutions in each step, whereas all-at once versions work with linear PDE (and adjoint) solves only. For the latter we refer to [1].

References

1 B. Kaltenbacher, A. Kirchner, B. Vexler: Goal oriented adaptivity in the IRGNM for parameter identification in PDEs II: all-at-once formulations. Inverse Problems 30 (2014) 045002.

3.15 Regularity of Model, Adjoint and its Relation to Optimization (Data Assimilation)

Peter Korn (MPI für Meteorologie – Hamburg, DE)

The problem of determining for a coupled set of nonlinear partial differential equations the initial condition from which a model trajectory emerges in agreement with a given set of time-distributed observations is studied by using a variational data assimilation approach. The partial differential equations describe a simplified coupled Atmospheric-Ocean model and consist of a coupled set of shallow-water equation in geophysical appropriate scaling. For the coupled model the existence of optimal initial conditions in the sense of minimizers of a specific cost functional and a first-order necessary condition involving the coupled adjoint equations are proven. Instrumental for the results are derivative based norms in the data assimilation cost functional such that Sobolev norms replace the standard Lebesgue norms.

References

1 Peter Korn. Fitting A Coupled Atmosphere-Ocean Model to Observations using Derivative-Based Norms, preprint, 2014

3.16 Constraint handling for gradient-based optimization of compositional reservoir flow

Drosos Kourounis (University of Lugano, CH)

Joint work of Kourounis, Drosos; Durlofsky, Louis; Jansen, Jan Dirk; Aziz, Khalid
Main reference D. Kourounis, L. J. Durlofsky, J. D. Jansen, K. Aziz, “Adjoint formulation and constraint handling for gradient-based optimization of compositional reservoir flow,” Journal of Computational Geosciences, 18(2):117–137, 2014.
URL http://dx.doi.org/10.1007/s10596-013-9385-8

The development of adjoint gradient-based optimization techniques for general compositional flow problems is much more challenging than for oil-water problems due to the increased complexity of the code and the underlying physics. An additional challenge is the treatment of non smooth constraints, an example of which is a maximum gas rate specification in injection or production wells, when the control variables are well bottom-hole pressures. Constraint handling through lumping is a popular and efficient approach. It introduces a smooth function that approximates the maximum of the specified constraints over the entire model or on a well-by-well basis. However, it inevitably restricts the possible solution paths the optimizer may follow preventing it to converge to feasible solutions exhibiting higher optimal values. A simpler way to force feasibility, when the constraints are upper and lower bounds on output quantities, is to satisfy these constraints in the forward model. This heuristic treatment has been demonstrated to be more efficient than lumping and at the same time it obtained better feasible optimal solutions for several models of increased complexity. In this work a new formal constraint handling approach is presented. Necessary modifications of the nonlinear solver used at every time step during the forward simulation are also discussed. All these constrained handling approaches are applied in a gradient-based optimization framework for exploring optimal CO2 injection strategies that enhance oil recovery for a realistic offshore field, the Norne field. This alternative approach increases the oil production twofold over even the best of its respective competitors.

3.17 Tool-Demo: Algorithmic Differentiation by Overloading in C++ using dco/c++

Johannes Lotz (RWTH Aachen University, DE)

Joint work of Lotz, Johannes; Leppkes, Klaus; Naumann, Uwe
URL https://www.stce.rwth-aachen.de/software/dco_cpp.html

Algorithmic Differentiation (AD) is a widespread technique for the automatic generation of discrete adjoint codes.

AD can be applied by a compiler as a source-to-source transformation or by making use of operator overloading techniques as a built-in language feature. The AD community agrees on the fact that both ways come with advantages and disadvantages. The main advantage of operator overloading is the out-of-the-box coverage of the complete programming language. The efficiency of the generated code on the other hand is its disadvantage, and simultaneously the advantage of a source-to-source compiler. The coupling of both techniques is apparent and required. Nevertheless, no automatic technical solution is available.

In contrast to that, it has not yet become apparent, if the discrete or continuous adjoint approach is preferable. The main difference between the two ways of deriving the adjoint (the dual) is that the discrete approach inherits all discretization methods from the original problem (the primal). This results in a dual implementation, which is equivalent to a line-by-line derivative of the primal implementation and can therefore be generated by AD techniques. The continuous adjoint approach on the other hand assumes validity of the mathematical equations and on that basis derives the sensitivity equations, which are then to be solved. Discretization decision are to be made again.

Taking both observations into account, a valuable overloading tool should have a flexible and extensible interface to couple not only compiler generated code with the overloading tool, but also hand-written code, eventually being continuous adjoint solutions.

dco/c++ features multiple ways of teaching the tool adjoint knowledge on different levels of abstraction to support the user in implementing the different couplings described above.

3.18 Adjoint Numerical Libraries

Viktor Mosenkis (RWTH Aachen University, DE)

Numerical libraries are often used by scientist while writing their codes. This allows them to write the code faster and concentrate on their research rather than spend time on implementing and testing numerical algorithms. Once it comes to adjoin the code the users of these libraries have to solve the problem of providing adjoint version of the numerical library routines. Using Algorithmic Differentiation (AD) may fail because the source code of the routine is not available while writing a continuous adjoint version of the routine as described in [1] is not an easy assignment and requires a deeper insight into the algorithms. And there is still the problem of testing the code. In any case the library supplier is the natural instance for providing adjoint version of his library routines.

The Numerical Algorithm Group delivers adjoint version of their Fortran and C Library routines. AD tool dco/c++ is used to adjoin these routines. Two interfaces are offered. One interface for direct and fast integration in dco/c++. The other one to be used without dco/c++

References

1 Uwe Naumann and Johannes Lotz and Klaus Leppkes and Markus Towara. *Algorithmic Differentiation of Numerical Methods: Tangent-Linear and Adjoint Solvers for Systems of Nonlinear Equations*. AIB-2012-15, Aachen, Germany, 2012

3.19 Moans about discrete adjoints

Jens-Dominik Mueller (Queen Mary University of London, GB)

Joint work of Mueller, Jens-Dominik; Shenren Xu; Marcus Meyer
Main reference S. Xu, D. Radford, M. Meyer, J.-D. Müller, "Stabilization of discrete steady adjoint solvers," submitted to JCP.

Discrete adjoints can very be produced very effectively with AD tools and promise to provide exact derivatives of the primal code. The former is essential for code maintenance and evolution: e. g. continuous adjoints typically are much less developed than their primal counterparts. The latter is relevant for advanced adjoint applications such as co-Kriging or uncertainty analysis. However there are pitfalls for discrete adjoints, two of which are highlighted.

For industrial CFD applications where the primal already is close to exhausting the available hardware, the preferred approach is the steady-state adjoints of the steady-state primals of industrial CFD applications. Quite frequently though an adjoint based on a fixed-point paradigm diverges since the primal converges only to limit cycle oscillations, as it is not contractive but its Jacobian possesses unstable eigenvalues. Using a stronger preconditioner can achieve convergence for primal and adjoint in cases of mild instability if used for both primal and adjoint.

On the other hand, counterexamples with continuous adjoints demonstrate that the added stabilization can help to converge the steady-state adjoint even on time-averaged unsteady primals. While a full unsteady approach with checkpointing of the primal will succeed for moderately chaotic flows, adding stabilization to discrete adjoints and quantifying this error should be considered as a robust and very cost-effective way to simulate highly turbulent flows, which in turn also may avoid issues of blow-up related to chaotic behavior.

Issues also arise with the use of point values of the discrete adjoint. Consistent continuous adjoint formulations converge to the analytic adjoint solution, simple examples of uniform channel flows with either Dirichlet or Neumann conditions demonstrate that the discrete adjoint based on a standard primal discretization does not. This in turn precludes the use of point values (rather than integrals) of the adjoint solution, or e. g. the use of boundary formulations of the sensitivity.

While finite-element discretizations naturally posses some dual consistency, for finite volume methods the implications of dual inconsistency and the possible gains offered by modified primals that provide dual-consistent discrete adjoints need more exploration.

3.20 Adjoint-Based Research and Applications at NASA Langley Research Center

Eric Nielsen (NASA Langley ASDC – Hampton, US)

Joint work of Nielsen, Eric; Diskin, Boris
Main reference E. Nielsen, B. Diskin, "Discrete Adjoint-Based Design for Unsteady Turbulent Flows on Dynamic Overset Unstructured Grids," AIAA Journal, 51(6):1355–1373, June 2013; available also from NASA repository.
URL http://dx.doi.org/10.2514/1.J051859
URL http://fun3d.larc.nasa.gov/papers/overset_adjoint_aiaaj.pdf

An overview of the use of adjoint methods at NASA Langley Research Center has been presented. Among major areas of large-scale adjoint-based applications are shape optimization, grid adaptation, and multidisciplinary optimization. In this talk, examples of unsteady adjoint-based aerodynamic shape optimization have been presented such as active flow control for a high-lift configuration, a helicopter in forward flight, a fighter jet with simulated aeroelastic effects, and a biologically-inspired flapping wing configuration. Several examples of adjoint-based mesh adaptation for various applications have included high-lift and nozzle plume configurations, a sonic boom application, an example of shock-boundary layer interaction, and several other aerospace applications. Multidisciplinary optimization capabilities have been demonstrated for sonic boom mitigation. High-performance computing aspects have also been discussed, such as scaling performance in the presence of frequent I/O for unsteady adjoint simulations.

3.21 Goal-oriented mesh adaptation based on total derivative of goal with respect volume mesh coordinates

Jacques Peter (ONERA – Châtillon, FR)

Main reference M. Nguyen-Dinh, J. Peter, R. Sauvage, M. Meaux, J.-A. Désidéri, "Mesh quality assessment based on aerodynamic functional output total derivatives," European Journal of Mechanics – B/Fluids, 45(May–June 2014):51–71, 2014.
URL http://dx.doi.org/10.1016/j.euromechflu.2013.12.004

In aeronautical CFD, engineers require accurate predictions of the forces and moments but they are less concerned with flow-field accuracy. Hence, the so-called "goal oriented" mesh adaptation strategies have been introduced to get satisfactory values of functional outputs at an acceptable cost, using local node displacement and insertion of new points rather than mesh refinement guided by uniform accuracy. Most often, such methods involve the adjoint vector of the function of interest. Our purpose is to present a new goal oriented criterion of mesh quality and a new local mesh adaptation strategy in the framework of finite-volume schemes and a discrete adjoint vector method. They are based on the total derivative of the goal with respect to mesh nodes coordinates. More precisely, a projection of the goal derivative, removing all components corresponding to geometrical changes in the solid walls or the support of the output is introduced. The norm of this vector field times the local characteristic mesh size is the proposed mesh adaptation criterion. The methods is assessed in the case of 2D and 3D Euler flow computations.

3.22 Numerical Algorithms for Mean-field type control Problems

Olivier Pironneau (UPMC – Paris, FR)

Main reference M. Lauriere, O. Pironneau, "Dynamic programming for mean-field type control," C.R.A.S Serie I, 1–7, Oct. 2014.
URL http://hal.archives-ouvertes.fr/hal-01018361

Mean-field type controls are stochastic optimization problems involving statistical functions of the state and/or control such as their means and variance. This happens often in problems modeling risk for banks and energy optimization because typically some variables of the problem depend on the mean behavior of all actors, each optimizing the same cost function at their individual level.

Stochastic control is best analyzed by Dynamic Programming (DP) leading to the Hamilton-Jacobi-Belmann equation for the remaining cost V(t), i. e. the optimization function from t to T, knowing that x=x_t, the state at t.

But here (see [1]) to apply DP we need to know the PDF of x_t not just its value at t and so the HJB equation contains derivatives with respect to measures.

We show in the conference that at the algorithmic level the extended DP is equivalent to standard calculus of variation applied to the deterministic problem derived from the stochastic one via the Fokker-Planck equation for the PDF of x_t. Thus the problem of finding the gradients and adjoints is solved.

We will illustrate the approach on 4 semi-academic mean-field type control problems

References

1 M. Lauriere and O. Pironneau: Dynamic programming for mean-field type control. C.R.A.S Serie I, 1–7, Oct. 2014.

3.23 The Edge Pushing Algorithm for Computing Sparse Hessians

Alex Pothen (Purdue University, US)

We have revisited the Edge Pushing Algorithm for computing Hessians, proposed by Gower and Mello in 2012. We derive the algorithm using the notion of live variables from data flow analysis in compiler theory, showing that the algorithm maintains an invariant about the adjoints and Hessian matrix elements it computes at each step. We have implemented the algorithm to achieve correctness and efficiency in the context of computing Hessian matrices with the ADOL-C library for Algorithmic Differentiation. We have incorporated pre-accumulation in the computation to reduce the execution time. We provide rigorous complexity bounds for the algorithms and report execution times for a collection of test problems, including a mesh optimization problem. The results show that the Edge Pushing algorithm can be faster than the currently used algorithms (that compute compressed Hessian matrices via graph coloring) for some problems by one or two orders of magnitude, while also using less memory. Our implementation is available as open-source software, and will be included in a future release of ADOL-C.

References

1 R. M. Gower and M. P. Mello. *A new framework for Hessian Automatic Differentiation*. Optimization Methods and Software, vol. 27, 2012.

3.24 Hybrid Adjoint Approaches to Industrial Hydrodynamics

Thomas Rung (TU Hamburg-Harburg, DE)

Joint work of Rung, Thomas; Arthur, Stück; Jörn, Kröger

The contribution reports the recent progress of the development of an adjoint Navier-Stokes method for incompressible flows [1] and its application to industrial hydrodynamics [2].

When attention is directed to gradient-based optimization, two approaches – the discrete and the continuous approach – are conceivable. The present research is derived from the continuous adjoint approach and aims at three crucial aspects for its industrial applicability, i. e. (a) accuracy and consistency, (b) numerical robustness and stability, (c) numerical efficiency and parallel performance.

The last aspect is a frequently addressed concern in conjunction with large-scale industrial applications and motivates the use of a continuous adjoint approach. As opposed to this, the first aspect is often identified as the origin of numerical problems and accuracy issues of the continuous adjoint methodology [3]. In order to support the duality between the discretised versions of the primal and the dual problem, a consistent discretization is derived for the building blocks of the adjoint system. The employed term-by-term strategy is based on the utilized primal, unstructured finite-volume discretization and their discrete adjoints obtained from the variation of the discrete Lagrangian. The strategy inheres continuous and discrete elements and is thus labeled hybrid adjoint approach. The approach supports the robustness of the algorithm and provides insight into an appropriate treatment of the adjoint coupling terms. Moreover, it facilitates a unified, discrete formulation of the adjoint wall-boundary condition and the related boundary-based sensitivity equation. Further details are found in [4].

References

1 A. Stück: Adjoint Navier-Stokes Methods for Hydrodynamic Shape Optimization. Dissertation, Hamburg University of Technology (TUHH), 2011.

2 A. Stück, J. Kröger, T. Rung: Adjoint-based hull design for wake optimization. Ship Technol. Res., 58(1):34-44, 2011.

3 J. E. V. Peter and R. P. Dwight. Numerical sensitivity analysis for aerodynamic optimization: A survey of approaches. Computers & Fluids, 39(3):373–391, 2010.

4 A. Stück, T. Rung. Adjoint complement to viscous finite-volume pressure-correction methods. Journal of Computational Physics, 248:402–419, 2013.

3.25 Structure Exploitation, AD and the Continuous Problem

Stephan Schmidt (Universität Würzburg, DE)

Hybridizing the discrete and continuous adjoint approach into a holistic, structure exploiting process is discussed. The potential advantages of this approach are exemplified within applications in the field of shape optimization as well as automatic code generation for FEM-problems such as provided by FEniCS.

Building upon previous studies [1], a potential hybrid Shape-AD tool could exploit the surface representation of the shape gradient, thereby automatically generating a primal/dual solver that can assemble a shape derivative by evaluating boundary quantities only. This could potentially lead to very fast and efficient code, automatically circumventing the necessity to consider mesh- or metric derivatives altogether by exploitation of the continuous problem structure.

Application examples where such a tool would be beneficial – starting by CFD problems and concluding with inverse design in acoustics and electromagnetism – are considered. The partial output of a preliminary semantic tool analyzing a shape optimization problem within the continuous Euler equations is shown as well.

References

1 C. Ilic, S. Schmidt, V. Schulz, and N. Gauger. Detailed aerodynamic shape optimization based on an adjoint method with shape derivatives. In M. Beckers, J. Lotz, V. Mosenkis, and U. Naumann, editors, *Fifth SIAM Workshop on Combinatorial Scientific Computing*, volume 2011-09, pages 8–10. Aachener Informatik-Berichte (AIB), 2011. ISSN 0935–3232.

3.26 Linking Adjoint Based Shape Optimization to Riemannian Geometry in Shape Spaces

Volker Schulz (Universität Trier, DE)

Main reference V. Schulz, "A Riemannian View on Shape Optimization," arXiv:1203.1493v2 [math.OC], 2014.
URL http://arxiv.org/abs/1203.1493v2

Shape optimization is a very active field of research with numerous applications of economic importance. Several examples from aerodynamics, acoustics and thermoelastics are used to illustrate this and to motivate the following more theoretical considerations. Although parametric geometry description (like CAD) are widely used in industry, they lead to high numerical costs for non-trivial geometry resolutions and pose severe limitations to the set of reachable shapes. The alternative avoiding these problems is the nonparametric approach which leaves all mesh nodes describing the geometry under investigation free for optimization and is based on the shape calculus. The current numerical state of the art in shape optimization based on the shape calculus is characterized by first order methods of steepest descent type and a general lack of second order methods. However, ideas from second order methods aiming a Newton-like strategies give rise to excellent preconditioners as demonstrated.

In this talk, a general framework based on differential geometric investigations is presented, which considers the set of admissible shapes as a Riemannian manifold and constructs Taylor series expansion and Newton methods similar to optimization ideas for finite dimensional matrix manifold. This novel approach, introduces a Riemannian shape Hessian as a Hessian formulation for second shape derivatives which, in contrast to the second shape derivative (which is so far historically but misleadingly named shape Hessian) possesses the properties which are expected from a Hessian: symmetry and provision of a Taylor series expansion. This approach is carried on to PDE constrained shape optimization and develops a novel sequential quadratic programming framework for shape optimization based on the shape calculus, where the linear-quadratic subproblems to be solved in each nonlinear iteration have the structure of usual optimal control problems and are thus accessible to the wealth of efficient methods developed for this problem class like, e. g., multigrid optimization methods.

3.27 Discrete Adjoint Optimization for OpenFOAM

Markus Towara (RWTH Aachen University, DE)

Joint work of Towara, Markus; Sen, Arindam; Naumann, Uwe
Main reference M. Towara, U. Naumann, "A Discrete Adjoint Model for OpenFOAM," in Proc. of the 2013 Int'l Conf. on Computational Science, Procedia Computer Science, Vol. 18, pp. 429–438, 2013.
URL http://dx.doi.org/10.1016/j.procs.2013.05.206

OpenFOAM is an Open-Soure CFD Simulation Tool with a wide range of applications and a strongly growing user base in both academia and industry. The source code is available in C++. The application of the adjoint model is a common approach for high dimensional optimization problems, however often a continuous approach instead of a discrete one is used. Codes which generate adjoint sensitivity information using the discrete approach are usually generated by Algorithmic Differentiation[1], either by source code transformation or operator overloading, thus differentiating the code on a per statement level instead of deriving and discretizing a set of adjoint equations as with the continuous approach. We introduced a discrete adjoint version of OpenFOAM using the operator overloading tool dco/c++ in order to generate derivatives and to apply optimization (with a focus on topology optimization)[2]. A discrete adjoint implementation in general yields a significant overhead to the passive evaluation in computation time and more importantly required memory. (Intermediate values from the whole computational history have to be stored in order to evaluate the derivatives). Our work focuses on how we managed to significantly reduce the memory requirements of said discrete adjoint OpenFOAM version, i. e. by applying analytical knowledge about the iterative linear solvers used to solve the underlying partial differential equations, thus eliminating the need to store intermediate values generated during the solution process of the linearized equations[3]. This treatment yields a significant improvement in both runtime and memory usage and also eliminates the dependency of the memory usage on the number of linear solver iterations.

References

1 A. Griewank and A. Walther, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, SIAM, 2008

2 M. Towara and U. Naumann, *A Discrete Adjoint Model for OpenFOAM*, Procedia Computer Science Vol. 18, 2013

3 M. Giles, *Collected Matrix Derivative Results for Forward and Reverse Mode Algorithmic Differentiation*, Advances in Automatic Differentiation, 2008

3.28 Convergence of discrete adjoints for flows with shocks

Stefan Ulbrich (TU Darmstadt, DE)

Joint work of Giles, Mike; Ulbrich, Stefan
Main reference M. Giles, S. Ulbrich, “Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 1: Linearized approximations and linearized output functionals,” SIAM J. on Numerical Analysis, 48(3):882–904, 2010.
URL http://dx.doi.org/10.1137/080727464
Main reference M. Giles, S. Ulbrich, “Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 2: Adjoint Approximations and Extensions,” SIAM J. on Numerical Analysis, 48(3):905–921, 2010.
URL http://dx.doi.org/10.1137/09078078X

We analyze the convergence of discrete adjoint approximations for optimal control problems governed by an unsteady one-dimensional hyperbolic conservation law with a convex flux function. A simple modified Lax-Friedrichs discretization is used on a uniform grid that has a numerical viscosity of $O(h^\alpha)$, $2/3 < \alpha < 1$, and we consider a tracking type objective function at the end time. The control are the initial data. It is known that such tracking type objective functions are differentiable with respect to the initial control also in the case of shocks and that an adjoint based representation of the reduced gradient of the objective can be obtained [1, 4, 5]. We show that the discrete adjoint scheme converges pointwise almost everywhere and uniformly outside of any neighborhood of the extreme backward characteristics emanating from shocks, see [2, 3]. A key point is that the numerical smoothing increases the number of points across the nonlinear discontinuity as the grid is refined. Hence, the discrete adjoint leads to a convergent representation of the reduced objective gradient. We sketch the proof idea of [2, 3] which is based on an asymptotic expansion with respect to the viscosity parameter in an inner region around the shock and an outer region. In addition, we present numerical results illustrating the asymptotic behavior which is analyzed. Finally, we illustrate that a numerical viscosity of $O(h^\alpha)$, $2/3 < \alpha < 1$ is necessary to obtain convergence of the discrete adjoint if a shock is present in the region, where the objective function is evaluated.

References

1 Michael B. Giles, *Discrete adjoint approximations with shocks*, in Hyperbolic Problems: Theory, Numerics, Applications, T. Hou and E. Tadmor, eds., Springer-Verlag, New York, 2003

2 Mike Giles and Stefan Ulbrich, *Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 1: Linearized approximations and linearized output functionals*, SIAM J. Numer. Anal., 48 (2010), pp. 882–904.

3 Mike Giles and Stefan Ulbrich, *Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 2: Adjoint approximations and extensions*, SIAM J. Numer. Anal., 48 (2010), pp. 905–921.

4 Stefan Ulbrich, *A sensitivity and adjoint calculus for discontinuous solutions of hyperbolic conservation laws with source terms*, SIAM J. Control Optim., 41 (2002), pp. 740–797

5 Stefan Ulbrich, *Adjoint-based derivative computations for the optimal control of dis- continuous solutions of hyperbolic conservation laws*, Systems Control Lett., 48 (2003), pp. 313–328.

3.29 White box adjoining of a radiative transport model and its use for ill-posed inverse problems

Joern Ungermann (Forschungszentrum Jülich, DE)

Main reference J. Ungermann, "Tomographic reconstruction of atmospheric volumes from infrared limb-imager measurements," Dissertation, Wuppertal University, Schriften des Forschungszentrums Jülich, Series Energie & Umwelt, Vol. 106, 153 pages, 2011.
URL http://hdl.handle.net/2128/4385

The Gimballed Limb Radiance Imager of the Atmosphere (GLORIA) is a newly developed unique atmospheric sounder that combines for the first time a classical Fourier transform spectrometer (FTS) with a 2-D detector array. Imaging allows the spatial sampling to be improved by up to an order of magnitude when compared to a conventional limb scanning instrument. GLORIA is designed to operate on various high altitude aircrafts.

Its unique scanning scheme and data acquisition rate allows for the first time the tomographic measurement of large air volumes about a thousand kilometers across. Reconstructing 3-D volumes from the measured infrared spectra thus poses a large-scale inverse problem, which requires a highly optimized forward model and inversion scheme.

Implementing an adjoint version using the dco tool suite of STCE, RWTH Aachen University, was the first step towards that goal. Separating the forward model internally into a linear map and a set of functions with one-dimensional output allowed to construct the full Jacobian matrix of the forward model by just a single execution without checkpointing. The performance could be further increased by a factor of two by manually computing the Jacobians of often-called subroutines and directly inserting these values into the tape, thereby drastically reducing tape-size.

Solving the inverse problem requires the minimization of a cost function composed of a term describing the agreement between measurements and simulated measurements for a given atmospheric state on the one hand and a regularizing term on the other hand. Having available the full Jacobian matrix of the forward model allows the efficient implementation of Quasi-Newton methods to minimize the cost function. These methods approximate the Hessian of the cost function by neglecting the Hessian of the forward model. The quadratic convergence of these methods requires usually less than 10 iterations for sufficient results and thereby only as many evaluations of the Jacobian matrix of the forward model. The numerical properties of the minimizer were greatly increased by providing an approximate Jacobian preconditioner for the Hessian of the cost function, which can be straightforwardly approximated given the available matrices. Lastly, each Quasi-Newton iteration requires the solution to a linear equation system, which can be produced matrix-free using conjugate-gradients. The regularizing properties of the conjugate-gradient scheme are used to implement a trust-region method, where intermediate solutions of increasing accuracy are stored along the computation of the precise solution and used in a back-tracking fashion to prevent the costly solving of additional linear equation systems.

3.30 Time-minimal Checkpointing

Andrea Walther (Universität Paderborn, DE)

Joint work of Walther, Andrea; Griewank, Andreas

For adjoint calculations, parameter estimation, and similar purposes one may need to reverse the execution of a computer program. The simplest option is to record a complete execution log and to read it backwards as required. This approach may require massive amounts of storage. Instead one may generate the execution log piecewise by restarting the "forward" calculation repeatedly from suitably placed checkpoints.

The basic structure of the resulting reversal schedules is illustrated. Various strategies are analyzed with respect to the resulting temporal and spatial complexity on serial and parallel machines. For serial machines known optimal compromises between operations count and memory requirement are explained.

For program execution reversal on multi-processors the new challenges and demands on an optimal reversal schedule are described. We present parallel reversal schedules that are provably optimal with regards to the number of concurrent processes and the total amount of memory required. More details on this time-minimal checkpointing approach can be found in [1, 2].

References

1 Andrea Walther: Program Reversal Schedules for Single- and Multi-processor Machines. Promotion, TU Dresden, 2000.

2 Andrea Walther: Bounding the number of processes and checkpoints needed in time-minimal parallel reversal schedules. Computing 731:35–154 (2004).

3.31 Adjoining Chaos

Qiqi Wang (MIT – Cambridge, US)

Joint work of Blonigan, Patrick; Gomez, St,even; Hu, Rui; Gauger, Nico; Diskin, Boris; Nielson, Eric; Farrell, Patrick
Main reference Q. Wang, R. Hu, P. Blonigan, "Least Squares Shadowing sensitivity analysis of chaotic limit cycle oscillations," Journal of Computational Physics, 267(June 2014):210–224, 2014; pre-print available as arXiv:1204.0159v7 [physics.comp-ph].
URL http://dx.doi.org/10.1016/j.jcp.2014.03.002
URL http://arxiv.org/abs/1204.0159v7

The adjoint method, among other sensitivity analysis methods, can fail in chaotic dynamical systems. The result from these methods can be too large, often by orders of magnitude, when the result is the derivative of a long time averaged quantity. This failure is known to be caused by ill-conditioned initial value problems. This paper overcomes this failure by replacing the initial value problem with the well-conditioned least squares shadowing (LSS) problem. The LSS problem is then linearized in our sensitivity analysis algorithm, which computes a derivative that converges to the derivative of the infinitely long time average. We demonstrate our algorithm in several dynamical systems exhibiting both periodic and chaotic oscillations.

Participants

- Christian Bischof
TU Darmstadt, DE
- Luise Blank
Universität Regensburg, DE
- David Bommes
INRIA Sophia Antipolis – Méditerranée, FR
- Martin Bücker
Universität Jena, DE
- Bruce Christianson
University of Hertfordshire, GB
- Mike Dewar
NAG Ltd. – Oxford, GB
- Boris Diskin
National Institute of Aerospace – Hampton, US
- Patrick Farrell
University of Oxford, GB
- Christian Frey
DLR – Köln, DE
- Nicolas R. Gauger
TU Kaiserslautern, DE
- Andreas Griewank
HU Berlin, DE
- Stefanie Günther
TU Kaiserslautern, DE
- Ralf Hannemann-Tamas
Univ. of Science & Technology – Trondheim, NO
- Laurent Hascoet
INRIA Sophia Antipolis – Méditerranée, FR
- Patrick Heimbach
MIT – Cambridge, US
- Paul D. Hovland
Argonne National Laboratory, US
- Barbara Kaltenbacher
Universität Klagenfurt, AT
- Peter Korn
MPI für Meteorologie – Hamburg, DE
- Drosos Kourounis
University of Lugano, CH
- Ben Lenser
HU Berlin, DE
- Johannes Lotz
RWTH Aachen, DE
- Marcus Meyer
Rolls-Royce – Blankenfelde-Mahlow, DE
- Viktor Mosenkis
RWTH Aachen, DE
- Jens-Dominik Mueller
Queen Mary University of London, GB
- Uwe Naumann
RWTH Aachen, DE
- Eric Nielsen
NASA Langley ASDC – Hampton, US
- Jacques Peter
ONERA – Châtillon, FR
- Olivier Pironneau
UPMC – Paris, FR
- Alex Pothen
Purdue University, US
- Thomas Rung
TU Hamburg-Harburg, DE
- Ekkehard Sachs
Universität Trier, DE
- Max Sagebaum
TU Kaiserslautern, DE
- Stephan Schmidt
Universität Würzburg, DE
- Volker Schulz
Universität Trier, DE
- Markus Towara
RWTH Aachen, DE
- Stefan Ulbrich
TU Darmstadt, DE
- Jörn Ungermann
Forschungszentrum Jülich, DE
- Jean Utke
Allstate – Northbrook, US
- Andrea Walther
Universität Paderborn, DE
- Qiqi Wang
MIT – Cambridge, US

Report from Dagstuhl Seminar 14372

Analysis of Algorithms Beyond the Worst Case

Edited by

Maria-Florina Balcan[1], Bodo Manthey[2], Heiko Röglin[3], and Tim Roughgarden[4]

1 **Carnegie Mellon University, US,** `ninamf@cs.cmu.edu`
2 **University of Twente, NL,** `b.manthey@utwente.nl`
3 **Universität Bonn, DE,** `roeglin@cs.uni-bonn.de`
4 **Stanford University, US,** `tim@cs.stanford.edu`

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 14372 "Analysis of Algorithms Beyond the Worst Case".

The theory of algorithms has traditionally focused on worst-case analysis. This focus has led to both a deep theory and many beautiful and useful algorithms. However, there are a number of important problems and algorithms for which worst-case analysis does not provide useful or empirically accurate results. This is due to the fact that worst-case inputs are often rather contrived and occur hardly ever in practical applications. Only in recent years a paradigm shift towards a more realistic and robust algorithmic theory has been initiated. The development of a more realistic theory hinges on finding models that measure the performance of an algorithm not only by its worst-case behavior but rather by its behavior on "typical" inputs. In this seminar, we discussed various recent theoretical models and results that go beyond worst-case analysis.

The seminar helped to consolidate the research and to foster collaborations among the researchers working in the different branches of analysis of algorithms beyond the worst case.

Seminar September 7–12, 2014 – http://www.dagstuhl.de/14372

1998 ACM Subject Classification F.2 Analysis of Algorithms and Problem Complexity, G.1.6 Optimization, G.2 Discrete Mathematics, G.3 Probability and Statistics, I.2.6 Learning, I.5.3 Clustering

Keywords and phrases analysis of algorithms, probabilistic analysis, smoothed analysis, approximation stability, machine learning

Digital Object Identifier 10.4230/DagRep.4.9.30

Edited in cooperation with Michael Etscheid

1 Executive Summary

Maria-Florina Balcan
Bodo Manthey
Heiko Röglin
Tim Roughgarden

The theory of algorithms has traditionally focused on worst-case analysis. This focus has led to both a deep theory and many beautiful and useful algorithms. There are, however, a number of important problems and algorithms for which worst-case analysis does not provide useful or empirically accurate results. For example, worst-case analysis suggests that the simplex method is an exponential-time algorithm for linear programming, while in fact it runs in near-linear time on almost all inputs of interest. Worst-case analysis ranks all deterministic caching algorithms equally, while in almost all applications some algorithms (like least-recently-used) are consistently superior to others (like first-in-first-out).

The problem is that worst-case analysis does not take into consideration that worst-case inputs are often rather contrived and occur hardly ever in practical applications. It led to the situation that for many problems the classical theory is not able to classify algorithms meaningfully according to their performance. Even worse, for some important problems it recommends algorithms that perform badly in practice over algorithms that work well in practice only because the artificial worst-case performance of the latter ones is bad.

Only in recent years a paradigm shift towards a more realistic and robust algorithmic theory has been initiated. The development of a more realistic theory hinges on finding models that measure the performance of an algorithm not only by its worst-case behavior but rather by its behavior on typical inputs. However, for an optimization problem at hand it is usually impossible to rigorously define the notion of "typical input" because what such an input looks like depends on the concrete application and on other indistinct parameters. The key to building a rigorous and realistic theory is hence not to define exactly what a typical input looks like, but to identify common properties that are shared by real-world inputs. As soon as such properties are identified, in many cases one can explain why certain heuristics work well in practice while others do not. The next step is then to look for algorithmic means to exploit these properties explicitly in order to obtain improved algorithms in practice that are not tailored to unrealistic worst-case inputs.

Many different models that go beyond classical worst-case theory have been suggested. These models can be divided into two main categories: either they are based on the assumption that inputs are to some extend random (probabilistic analysis) or they consider only inputs that satisfy certain deterministic properties that mitigate the worst case.

Probabilistic Analysis

Average-case analysis is probably the first thought that springs to mind when mentioning probabilistic input models. In such an analysis, one considers the expected performance on random inputs. Starting in the seventies, many algorithms that showed a remarkable performance in practice have been analyzed successfully on random inputs. This includes algorithms for classical optimization problems, such as the traveling salesman problem, or the simplex method for linear programming.

While average-case analysis has been successfully applied to many problems, a major concern is that inputs chosen completely at random have for many problems little in common with inputs arising in practice. Similar as a random TV screen produced by static noise has nothing to do with a typical TV screen, a set of random points does not resemble, for instance, a realistic clustering instance with mostly well-separated clusters.

Smoothed Analysis. To overcome the drawbacks of average-case and worst-case analysis, the notion of smoothed analysis has been suggested by Spielman and Teng in 2001. In this model, inputs are generated in two steps: first, an adversary chooses an arbitrary instance, and then this instance is slightly perturbed at random. The smoothed performance of an algorithm is defined to be the worst expected performance the adversary can achieve. This model can be viewed as a less pessimistic worst-case analysis, in which the randomness rules out pathological worst-case instances that are rarely observed in practice but dominate the worst-case analysis. If the smoothed running time of an algorithm is low (i. e., the algorithm is efficient in expectation on any perturbed instance) and inputs are subject to a small amount of random noise, then it is unlikely to encounter an instance on which the algorithm performs poorly. In practice, random noise can stem from measurement errors, numerical imprecision or rounding errors. It can also model arbitrary influences, which we cannot quantify exactly, but for which there is also no reason to believe that they are adversarial.

The framework of smoothed analysis was originally invented to explain the practical success of the simplex method for linear programming. Spielman and Teng analyzed linear programs in which each coefficient is perturbed by Gaussian noise with standard deviation σ. They showed that the smoothed running time of the simplex method is bounded polynomially in the input size and $1/\sigma$. Hence, even if the amount of randomness is small, the expected running time of the simplex method is polynomially bounded. After its invention smoothed analysis has attracted a great deal of attention and it has been applied in a variety of different contexts, e. g., in multi-objective optimization, local search, clustering, and online algorithms. By now smoothed analysis is widely accepted as a realistic alternative to worst-case analysis.

Semi-random Models. Semi-random input models can be considered as analogues of smoothed analysis for graph problems and they even predate smoothed analysis by a couple of years. There is a variety of semi-random graph models that go beyond the classical Erdős-Rényi random graphs. In most of these models graphs are generated by a noisy adversary – an adversary whose decisions (whether or not to insert a particular edge) have some small probability of being reversed. Another well-studied class of semi-random models are *planted models*, in which a solution (e. g., an independent set or a partitioning of the vertices in color classes) is chosen and then edges are added randomly or by an adversary of limited power in such a way that the given solution stays a valid solution for the given problem.

Similar to smoothed analysis, semi-random models have been invented in order to better understand the complexity of NP-hard graph problems because Erdős-Rényi random graphs often do not reflect the instances one encounters in practice – many graph problems are quite easy to solve on such random graphs.

Deterministic Input Models

Smoothed analysis and semi-random models are multi-purpose frameworks that do not require much information about how exactly typical inputs for the optimization problem at hand look like. If more information is available, it makes sense to identify structural properties of typical inputs that make them easier to solve than general inputs. There are well known examples of this approach like the TSP, which gets easier (in terms of approximation) when restricted to inputs in which the distances satisfy the triangle inequality. Also in computational geometry it is a very common phenomenon that problems become easier if one assumes that no angles are too small or not too many objects overlap in the same region.

In recent years there has been an increased interest in more sophisticated deterministic input models, in particular for clustering problems. Balcan, Blum, and Gupta introduce and exploit the so-called $(1 + \alpha, \varepsilon)$-approximation-stability property of data in the context of clustering. This assumption is motivated by the observation that in many clustering applications there is usually one correct but unknown target clustering, and the goal is to find a clustering that is close to this target clustering and misclassifies only a few objects. On the other hand in the common mathematical formulation of clustering problems a potential function is defined that assigns a value to each clustering. Then a clustering is computed that approximately optimizes the potential function (exact optimization is usually NP-hard). This approach makes sense only if clusterings that approximately optimize the potential function are close to the target clustering. Hence, an implicit assumption underlying this approach is that every clustering that approximately optimizes the objective function is close to the desired target clustering. Balcan et al. made this assumption explicit: they define that a clustering instance satisfies the $(1+\alpha, \varepsilon)$-approximation-stability assumption if in every c-approximation of the potential function at most an ε-fraction of all objects is misclassified compared to the target clustering. Balcan et al. showed that clustering instances with this property are easier to solve than general instances. They have shown specifically how to get ε-close to the target even for values of α for which finding a $1+\alpha$ approximation is NP-hard. Voevodoski et al. have shown that this approach leads to very efficient and accurate algorithms (with improved performance over previous state-of-the-art algorithms) for clustering biological datasets.

Bilu and Linial and later Awasthi, Blum, and Sheffet have considered instances of clustering problems that are *perturbation resilient* in the sense that small perturbations of the metric space do not change the optimal solution. They argue that interesting instances of clustering problems are stable and they prove that the assumption of stability renders clustering polynomial-time solvable. Balcan and Liang further relaxed this assumption to require only that the optimal solution after the perturbations is close to the optimal solution for the unperturbed instance.

These results have triggered a significant amount of work in the past years in the context of clustering and machine learning problems more generally, including subsequent works that proposed new related stability conditions (e. g., the “proximity condition” by Kannan and Kumar). Such works are very good examples demonstrating that identifying properties of real-world inputs can be extremely beneficial for our understanding of algorithmic problems.

Program of the Seminar

The program of the seminar consisted of 23 talks, including the following survey talks:

- Preprocessing of NP-hard problems, Uriel Feige;
- Approximation-stability and Perturbation-stability, Avrim Blum;
- Computational Feasibility of Clustering under Clusterability Assumptions, Shai Ben-David;
- Parametrizing the easiness of machine learning problems, Sanjoy Dasgupta;
- Linear Algebra++: Adventures and Unsupervised Learning, Sanjeev Arora.

The rest of the talks were 30-minute presentations on recent research of the participants. The time between lunch and the afternoon coffee break was mostly left open for individual discussions and collaborations in small groups. One open-problem session was organized.

One of the main goals of the seminar was to foster collaborations among the researchers working in the different branches of analysis of algorithms as sketched above. This is particularly important because at the moment the two communities dealing with probabilistic analysis and deterministic input models are largely disjoint. The feedback provided by the participants shows that the goals of the seminar, namely to circulate new ideas and create new collaborations, were met to a large extent.

The organizers and participants wish to thank the staff and the management of Schloss Dagstuhl for their assistance and support in the arrangement of a very successful and productive event.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 New Measures and Techniques in the Analysis of Online Search Problems

Spyros Angelopoulos (UPMC – Paris, FR)

Joint work of Angelopoulos, Spyros; Durr, Christoph; Lopez-Ortiz, Alejandro; Lucarelli, Giorgio; Panagiotou, Konstantinos

We consider the problem of exploring a set of concurrent rays using a single searcher. The rays are disjoint with the exception of a single common point, and in each ray a potential target may be located. The objective is to design efficient search strategies for locating the targets as quickly as possible.

In this talk we will describe some recent measures for the analysis of online search problems and in particular for the star-search problem. The new measures transcend the traditional worst-case analysis and give rise to new algorithmic approaches.

3.2 Adventures in Linear Algebra++ and Unsupervised Learning

Sanjeev Arora (Princeton University, US)

Linear Algebra++ is my name for some extensions to Linear Algebra problems (solving linear equations, rank, finding better basis to represent the data, eigenvalues/eigenvectors) that involve constraints such as sparsity, nonnegativity, and approximation (via ℓ_2 or another norm).

These variants arise in machine learning settings and are often NP-hard, i. e., difficult on worst-case inputs. Hence the only way to design provable algorithms for them is to go beyond worst case analysis.

The talk surveyed some recent work of mine and others that does this, with applications to ML problems such as topic modeling, sparse coding, nonnegative matrix factorization, deep learning, etc.

3.3 Learning Submodular Functions: An Analysis beyond the Worst Case

Maria-Florina Balcan (Carnegie Mellon University – Pittsburgh, US)

Joint work of Balcan, Maria-Florina; Harvey, Nicholas J. A.; Florin Constantin, Satoru Iwata, Lei Wang; Nan Du, Yingyu Liang, and Le Song

Main reference M.-F. Balcan, N. J. A. Harvey, "Learning Submodular Functions," in Proc. of the 43rd Annual ACM Symp. on Theory of Computing (STOC'11), pp. 793–802, ACM, 2011; pre-print available from author's webpage.

URL http://dx.doi.org/10.1145/1993636.1993741

URL http://www.cs.cmu.edu/~ninamf/papers/LearningSubmodular-STOC.pdf

Submodular functions are discrete functions that model laws of diminishing returns and enjoy numerous algorithmic applications that have been used in many areas, including social networks, machine learning, and economics. In this work we use a learning theoretic angle for studying submodular functions. We provide algorithms for learning submodular functions, as well as lower bounds on their learnability. In doing so, we uncover several novel structural results revealing both extremal properties as well as regularities of submodular functions, of interest to many areas.

Our lower bounds highlight the importance of providing an analysis on the learnability of these functions beyond the worst case. For classes of functions that exhibit additional structure in addition to diminishing returns (including probabilistic coverage functions and XOS functions with bounded complexity) we provide stronger guarantees on their learnability together with an interesting application for learning the influence function in a social network.

References

1 Maria Florina Balcan, Nick Harvey. *Learning submodular functions.* STOC 2011.

2 Maria Florina Balcan, Florin Constantin, Satoru Iwata, and Lei Wang. *Learning valuation functions.* COLT 2012.

3 Maria Florina Balcan, Nan Du, Yingyu Liang, and Le Song. *Influence Function Learning in Information Diffusion Networks.* ICML 2014.

3.4 Computational Efficiency of Clustering Well-Behaved Input Instances

Shai Ben-David (University of Waterloo, CA)

The goal of this talk is two-fold. First, I would like to provide a personally biased overview of the research concerning the computational complexity of clustering under data niceness assumptions. Having worked in this area for quite some time now, I feel that while the TCS community appreciates work that may have practical relevance (and clustering is clearly a task that arises in many applications), sometimes in this area there is a significant gap between research motivation and the actual technical results it yields. A secondary aim of this paper is to call the attention of the theoretical research community to some such gaps and encourage further work along directions that might otherwise have seemed resolved.

Computational complexity theory aims to provide tools for the quantification and analysis of the computational resources needed for algorithms to perform computational tasks. Worst-case complexity is by far the best known, most researched and best understood approach to computational complexity theory. In particular, NP-hardness is a worst-case-instance notion. By saying that a task is NP-hard (and assuming P $\neq$ NP), we imply that for every algorithm, there exist infinitely many instances on which it will have to work hard. However, for many problems this measure is unrealistically pessimistic compared to the experience of solving them for practical instances. A problem may be NP-hard and still have algorithms that solve it efficiently for any instance that is likely to occur in practice or any instance for which one cares about finding an optimal solution to.

Several approaches have been proposed to bringing computational complexity theory closer to the actual hardness faced when solving optimization problems on real data. *Average Case Complexity* ([8], [3]), analyzes run time w.r.t. some given probability distribution over the input instances. *Smoothed Analysis* ([9]) examines the running time of a given algorithm by taking the worst case over all inputs of the average runtime of the algorithm over some vicinity of the input. A different approach is to have a notion of "well-behaved-instances", so that on one hand it is reasonable to expect that instances one comes across in applications are so well behaved, and on the other hand there exist algorithms that can solve any well behaved input in polynomial time. Various earlier approaches have addressed computational hardness by defining subsets of relatively-easy instances (most notably, the area of *parameterized complexity* ([7])). [2] and [5] propose general notions of tamed instances that apply across different problems. Both of these papers apply some type of *robustness to perturbations* as the key property of such well behaved instances. Algorithms that efficiently solve NP-hard problems on such perturbation robust instances have been shown to exist for agnostic learning of half-spaces ([4]) and for graph partitioning problems ([6]). [1] formalized a *uniqueness of the optimal solution* criterion as a notion of well behaved clustering instances, which can also be applied to other types of problems. In this note I address the application of such approaches to clustering. I discuss those, as well as other notions of niceness-of-instances that are specific to clustering problems, as a basis for alternatives to worst-case complexity analysis of clustering tasks.

Clustering is a very useful paradigm that is being applied in a wide range of data exploration tasks. The term "clustering" should be thought of as an umbrella notion for a big and varied collection of tasks and algorithmic paradigms. Here, I will focus on clustering tasks that are defined as discrete optimization problems. Most of those optimization problems are NP-hard. I wish to examine whether this hardness remains an issue when we restrict our attention to "clusterable data" – data for which a meaningful clustering exists (one can argue that when there is no cluster structure in a given data set, there is no point in applying a clustering algorithm to it). In other words, we wish to evaluate to what extent current theoretical work supports the "Clustering is difficult only when it does not matter" (CDNM) thesis. A hint to my conclusion – the answer is a resounding NO.

I'll upload the full paper to arXiv soon.

References

1 Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In *Proc. of the 20th Annual ACM-SIAM Symp. on Discrete Algorithms, SODA 2009, New York, NY, USA, January 4–6, 2009*, pages 1068–1077, 2009.

2 Shai Ben-David. Alternative measures of computational complexity with applications to agnostic learning. In *Theory and applications of models of computation*, volume 3959 of *Lecture Notes in Comput. Sci.*, pages 231–235. Springer, Berlin, 2006.

3 Shai Ben-David, Benny Chor, Oded Goldreich, and Michael Luby. On the theory of average case complexity (abstract). In *Proc. of the 4th Annual Structure in Complexity Theory Conf., University of Oregon, Eugene, Oregon, USA, June 19–22, 1989*, page 36, 1989.

4 Shai Ben-David and Hans-Ulrich Simon. Efficient learning of linear perceptrons. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 189–195, 2000.

5 Yonatan Bilu and Nathan Linial. Are stable instances easy? In *Innovations in Computer Science – ICS 2010, Tsinghua University, Beijing, China, January 5–7, 2010. Proceedings*, pages 332–341, 2010.

6 Yonatan Bilu and Nathan Linial. Are stable instances easy? *Combinatorics, Probability & Computing*, 21(5):643–660, 2012.

7 R. G. Downey and M. R. Fellows. *Parameterized complexity.* Monographs in Computer Science. Springer-Verlag, New York, 1999.

8 Leonid A. Levin. Average case complete problems. *SIAM J. Comput.*, 15(1):285–286, 1986.

9 Daniel Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time. In *Proc. of the 33rd Annual ACM Symp. on Theory of Computing*, pages 296–305 (electronic). ACM, New York, 2001.

3.5 Approximation and Perturbation Stability

Avrim Blum (Carnegie Mellon University – Pittsburgh, US)

Joint work of Blum, Avrim; Balcan, Maria-Florina; Gupta, Anupam; Awasthi, Pranjal; Sheffet, Or; Vempala, Santosh

Often in optimization problems, the formal objective is a proxy for a goal of finding a desired "target" answer. For example, in clustering news articles by topic, one would like to find a clustering that agrees with how a human user would cluster them, or in segmenting a picture into objects, one would like to find a segmentation that agrees with the ground truth. If one encodes such a problem as (for instance) a k-means optimization problem, then implicitly one is hoping that the optimal solution to the objective corresponds to a solution that is similar to the desired target answer. This talk describes and discusses approximation-stability and perturbation-resilience, two stability notions motivated by making this and associated assumptions explicit. We present results for instances satisfying these conditions, focusing on clustering but also discussing Nash equilibria and related problems.

Nearly all of this work is joint with Nina Balcan. Major portions were also joint with Pranjal Awasthi, Anupam Gupta, Or Sheffet, and Santosh Vempala. Portions of the talk also discuss work that I was not involved in (e. g., of Nina Balcan and Bruce Liang).

3.6 Fréchet Distance: Synergies of Algorithms and Lower Bounds

Karl Bringmann (ETH Zürich, CH)

Joint work of Bringmann, Karl; Künnemann, Marvin
Main reference K. Bringmann, "Why walking the dog takes time: Fréchet distance has no strongly subquadratic algorithms unless SETH fails," in Proc. of the 55th IEEE Annual Symp. on Foundations of Computer Science (FOCS'14), pp. 661–670, IEEE CS, 2014; pre-print available as arXiv:1404.1448v2 [cs.CG].
URL http://dx.doi.org/10.1109/FOCS.2014.76
URL http://arxiv.org/abs/1404.1448v2

The Fréchet distance is a well-studied and very popular measure of similarity of two curves. Regarding worst-case complexity, all known algorithms to compute the Fréchet distance of two polygonal curves with n vertices have a runtime of $\tilde{\Theta}(n^2)$, omitting logarithmic factors. To obtain a conditional lower bound, we assume the Strong Exponential Time Hypothesis. Under this assumption we show that the Fréchet distance cannot be computed in strongly subquadratic time, i. e., in time $O(n^{2-\delta})$ for any $\delta > 0$. This means that finding faster algorithms for the Fréchet distance is as hard as finding faster algorithms for the satisfiability problem, and the existence of a strongly subquadratic algorithm can be considered unlikely.

To overcome the worst-case quadratic time barrier, restricted classes of curves have been studied that attempt to capture realistic input curves. The most popular such class are c-packed curves, for which the Fréchet distance has a $(1+\varepsilon)$-approximation in time $\tilde{O}(cn/\varepsilon)$ by Driemel, Har-Peled, and Wenk [3]. In dimension $d \geq 5$ we show that this cannot be improved to $O((cn/\sqrt{\varepsilon})^{1-\delta})$ for any $\delta > 0$ unless the Strong Exponential Time Hypothesis fails. Moreover, exploiting properties that prevent stronger lower bounds, we present an improved algorithm with runtime $\tilde{O}(cn/\sqrt{\varepsilon})$, matching our conditional lower bound.

This is partly based on joint work with Marvin Künnemann [2].

References

1 K. Bringmann. *Why walking the dog takes time: Fréchet distance has no strongly subquadratic algorithms unless SETH fails.* In: Proc. 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS'14), pp. 661–670, 2014.

2 K. Bringmann and M. Künnemann. *Improved approximation for Fréchet distance on c-packed curves matching conditional lower bounds.* Submitted. 2014. arXiv: 1408.1340.

3 A. Driemel, S. Har-Peled, and C. Wenk. *Approximating the Fréchet distance for realistic curves in near linear time.* Discrete & Computational Geometry 48.1 (2012), 95–127.

3.7 Smoothed Analysis of the Successive Shortest Path and Minimum-Mean Cycle Canceling Algorithms

Kamiel Cornelissen (University of Twente, NL)

Joint work of Brunsch, Tobias; Cornelissen, Kamiel; Manthey, Bodo; Röglin, Heiko
Main reference T. Brunsch, K. Cornelissen, B. Manthey, H. Röglin, "Smoothed Analysis of the Successive Shortest Path Algorithm," in Proc. of the 24th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA'13), pp. 1180–1189, SIAM, 2013.
URL http://dx.doi.org/10.1137/1.9781611973105.85

For minimum-cost maximum flow (MCF) algorithms, worst-case running time bounds do not always give a good indication for how the algorithms perform in practice. To model the performance on practical instances, we analyze two MCF algorithms, the successive shortest path (SSP) algorithm and the minimum-mean cycle canceling (MMCC) algorithm, in the framework of smoothed analysis. An adversary can specify the flow network and the edge capacities, but for the edge costs c_e the adversary can only specify density functions f_e of maximum density ϕ, according to which the edge costs are drawn.

We show that in the smoothed setting the SSP algorithm only needs $O(mn\phi(m + n\log n))$ running time, in stark contrast to the worst case running time. We also show almost tight lower bounds.

For the MMCC algorithm we show a smoothed running time of $O(m^2n^2(n\log n + \log\phi))$, which is better than the worst-case running time for dense graphs. We also show a lower bound of $\Omega(m^2n\log\phi)$.

3.8 The (Parameterized) Complexity of Counting Subgraphs

Radu Curticapean (Universität des Saarlandes, DE)

Joint work of Curticapean, Radu; Marx, Dániel
Main reference R. Curticapean, D. Marx, "Complexity of counting subgraphs: only the boundedness of the vertex-cover number counts," in Proc. of the 55th IEEE Annual Symp. on Foundations of Computer Science (FOCS'14), pp. 130–139, IEEE CS, 2014; pre-print available at arXiv:1407.2929v1 [cs.CC].
URL http://dx.doi.org/10.1109/FOCS.2014.22
URL http://arxiv.org/abs/1407.2929v1

We study the parameterized complexity of the following counting version of subgraph isomorphism: Given a "pattern" graph H and a "host" graph G, count the (not necessarily induced) copies of H in G, that is, subgraphs of G that are isomorphic to H.

In the first part, we follow an approach introduced by Marx and Pilipczuk, who studied the decision version of this problem under a set of 10 relevant parameters that can be imposed upon the graphs H and G, e. g., the size of the graphs, the number of connected components, maximum degree, treewidth, genus, and others. For each choice of parameters $k_1, \ldots, k_r$ and $e_1, \ldots, e_s$, they either gave an algorithm that solves subgraph isomorphism in time $f(k_1, \ldots, k_r) \cdot n^{g(e_1, \ldots, e_s)}$ for computable functions f, g – or they showed that such an algorithm would imply P = NP or FPT = W[1].

We are currently transferring this framework to the counting world and ask which parameter combinations admit such algorithms for the problem of counting subgraphs. Note that the set of algorithms can only shrink as counting is always harder than deciding. While this is still work in progress, we have almost obtained a dichotomy for the problem. In particular, we have identified some cases when counting is indeed harder than deciding.

In the second part, we consider only the parameter $|H|$ and ask for $f(|H|) \cdot n^{O(1)}$ algorithms for counting subgraphs. However, in this second problem, the pattern H may only be drawn from a fixed class C of graphs, which gives rise to the problem #Sub(C). There are very simple pattern classes C for which #Sub(C) is already known to be #W[1]-complete, such as the classes of paths, matchings or cycles. On the class of stars however, #Sub(C) actually even admits an $n^O(1)$ algorithm (in particular, this algorithm does not exploit the parameter $|H|$). We want to know for which classes #Sub(C) is fixed-parameter tractable (or even polynomial-time solvable).

It turns out that the only restriction that makes #Sub(C) easy is a bound on the maximum matching size in C. If C admits such a bound c, then #Sub(C) even admits a polynomial-time algorithm with an exponent depending only on c. If C however has unbounded matching size, then #Sub(C) is #W[1]-complete. In particular, we obtain no classes C for which #Sub(C) is FPT, but (probably) not in P, and thus, under the widely-believed assumption that FPT is not equal to #W[1], our dichotomy shows precisely which problems #Sub(C) admit polynomial-time algorithms.

3.9 Parametrizing the Easiness of Machine Learning Problems

Sanjoy Dasgupta (University of California – San Diego, US)

Joint work of Dasgupta, Sanjoy; Chaudhuri, Kamalika; Freund, Yoav; Kpotufe, Samory; Sinha, Kaushik

We consider a variety of basic machine learning tasks whose worst case complexity (computational, or statistical, or both) is exponentially bad. These include:

- Nonparametric classification and regression
- Nearest neighbor search and classification
- Density-based clustering

In each case, the lower bounds are based on constructions that may be considered pathological relative to instances that occur in practice. It has been fruitful to formally describe types of structure that occur in natural instances, and to find algorithms that are able to exploit such structure.

This undertaking is in its infancy, and there are many immediate open problems.

3.10 How to Serve Impatient Users

Matthias Englert (University of Warwick, GB)

Joint work of Cygan, Marek; Englert, Matthias; Gupta, Anupam; Mucha, Marcin; Sankowski, Piotr
Main reference M. Cygan, M. Englert, A. Gupta, M. Mucha, P. Sankowski, "Catch Them if You Can: How to Serve Impatient Users," in Proc. of the 4th Conf. on Innovations in Theoretical Computer Science (ITCS'13), pp. 485–494, 2013.
URL http://dx.doi.org/10.1145/2422436.2422489

Consider the following problem of serving impatient users: we are given a set of customers we would like to serve. We can serve at most one customer in each time step (getting value v_i for serving customer i). At the end of each time step, each as-yet-unserved customer i leaves the system independently with probability q_i, never to return. What strategy should we use to serve customers to maximize the expected value collected?

The standard model of competitive analysis can be applied to this problem: picking the customer with maximum value gives us half the value obtained by the optimal algorithm, and using a vertex weighted online matching algorithm gives us $1 - 1/e \approx 0.632$ fraction of the optimum. As is usual in competitive analysis, these approximations compare to the best value achievable by an clairvoyant adversary that knows all the coin tosses of the customers. Can we do better?

We show an upper bound of ≈ 0.648 if we compare our performance to such an clairvoyant algorithm, suggesting we cannot improve our performance substantially. However, these are pessimistic comparisons to a much stronger adversary: what if we compare ourselves to the optimal strategy for this problem, which does not have an unfair advantage? In this case, we can do much better: in particular, we give an algorithm whose expected value is at least 0.7 of that achievable by the optimal algorithm. This improvement is achieved via a novel rounding algorithm, and a non-local analysis.

3.11 Smoothed Analysis of Local Search for the Maximum-Cut Problem

Michael Etscheid (Universität Bonn, DE)

Joint work of Etscheid, Michael; Röglin, Heiko
Main reference M. Etscheid, H. Röglin, "Smoothed Analysis of Local Search for the Maximum-Cut Problem," in Proc. of the 25th ACM-SIAM Symp. on Discrete Algorithms (SODA'14), pp. 882–889, SIAM, 2014.
URL http://dx.doi.org/10.1137/1.9781611973402.66

Even though local search heuristics are the method of choice in practice for many well-studied optimization problems, most of them behave poorly in the worst case. This is in particular the case for the Maximum-Cut Problem, for which local search can take an exponential number of steps to terminate and the problem of computing a local optimum is PLS-complete. To narrow the gap between theory and practice, we study local search for the Maximum-Cut Problem in the framework of smoothed analysis in which inputs are subject to a small amount of random noise. We show that the smoothed number of iterations is quasi-polynomial, i. e., it is bounded from above by a polynomial in $n^{\log(n)}$ and ϕ where n denotes the number of nodes and ϕ denotes the perturbation parameter. This shows that worst-case instances are fragile and it is a first step in explaining why they are rarely observed in practice.

3.12 Preprocessing of NP-hard Problems

Uriel Feige (Weizmann Institute, IL)

Joint work of Feige, Uriel; Jozeph, Shlomo
Main reference U. Feige, S. Jozeph, "Universal Factor Graphs," in Proc. of the 39th Int'l Colloquium on Automata, Languages, and Programming (ICALP'12) – Part I, LNCS, Vol. 7931, pp. 339–350, Springer, 2012; pre-print available at arXiv:1204.6484v1 [cs.CC].
URL http://dx.doi.org/10.1007/978-3-642-31594-7_29
URL http://arxiv.org/abs/1204.6484v1

Consider a setting in which an input instance for an NP-hard optimization problem is supplied in two steps. In the first step, one gets to see some partial information about the input instance, referred to as a "preview". Based on this preview, a preprocessing algorithm can spend arbitrary (e. g., exponential) time in preparing some polynomial size "advice" string. In the second step, one gets to see the full input instance. Thereafter, a polynomial time algorithm attempts to solve the instance, and may use for this purpose the advice string prepared by the preprocessing algorithm. For various NP-hard optimization problems we present natural preview functions whose study appears to be well motivated. In certain cases we can prove that preprocessing leads to improved approximation ratios, and in certain cases we can prove limitations on how much preprocessing can help. Many natural questions within this framework remain open, and can serve as fertile ground for future research.

References

1 Uriel Feige, Shlomo Jozeph: *Universal Factor Graphs*. ICALP(1) 2012: 339–350
2 Uriel Feige, Shlomo Jozeph: *Demand Queries with Preprocessing*. ICALP(1) 2014: 477–488.

3.13 Average-Case Analysis of Parameterized Problems

Tobias Friedrich (Universität Jena, DE)

Joint work of Friedrich, Tobias; Nikolaos Fountoulakis; Danny Hermelin; Anton Krohmer

Many real world problems are NP-hard and are therefore generally believed not to be solvable in polynomial time. Additional assumptions on the inputs are necessary to solve such problems efficiently. Two typical approaches are (i) parameterized complexity where we assume that a certain parameter of the instances is small, and (ii) average-case complexity where we assume a certain probability distribution on the inputs. There is a vast literature on both approaches, but none about their intersection. Nevertheless, combining these two approaches seems natural and potentially useful in practice.

Analogous to the classical average-case complexity classes of Levin, we define a hierarchy of parameterized average-case complexity classes. To show the applicability of this theory we prove that the fundamental W[1]-complete problem k-Clique drops to the average-case analog of FPT for Erdős- Rényi random graphs of all densities and scale-free inhomogeneous random graphs.

3.14 On the Smoothed Approximation Performance of the 2-OPT Heuristic

Marvin Künnemann (MPI für Informatik – Saarbrücken, DE)

Joint work of Künnemann, Marvin; Manthey, Bodo

While the 2-OPT heuristic is a very simple and popular approach to solve Euclidean TSP, the theoretical understanding of its practical performance is still quite restricted. As a welcome exception, the smoothed analysis perspective proved useful to explain particularly the heuristic's fast running time observed in practice [1].

In this talk, we focus on the question of how well this paradigm also explains the practical approximation performance of the 2-OPT heuristic. We present new upper and lower bounds and, along the way, discuss aspects of the Gaussian and the ϕ-bounded perturbation models.

This is joint work (in progress) with Bodo Manthey.

References

1 Matthias Englert, Heiko Röglin, and Berthold Vöcking. Worst case and probabilistic analysis of the 2-opt algorithm for the TSP (extended abstract). In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1295–1304. ACM, New York, 2007.

3.15 Constant Factor Approximation for Balanced Cut in the PIE Model

Konstantin Makarychev (Microsoft Corp. – Redmond, US)

Joint work of Makarychev, Konstantin; Makarychev, Yury; Vijayaraghavan, Aravindan
Main reference K. Makarychev, Y. Makarychev, A. Vijayaraghavan, "Constant factor approximation for balanced cut in the PIE model." in Proc. of the 46th Annual ACM Symp. on Theory of Computing (STOC'14), pp. 41–49, ACM, 2014; pre-print available at arXiv:1406.5665v1 [cs.DS].
URL http://dx.doi.org/10.1145/2591796.2591841
URL http://arxiv.org/abs/1406.5665v1

We propose and study a new semi-random semi-adversarial model for Balanced Cut, a planted model with permutation-invariant random edges (PIE). Our model is much more general than planted models considered previously. Consider a set of vertices V partitioned into two clusters L and R of equal size. Let G be an arbitrary graph on V with no edges between L and R. Let E_{random} be a set of edges sampled from an arbitrary permutation-invariant distribution (a distribution that is invariant under permutation of vertices in L and in R). Then we say that $G + E_{\text{random}}$ is a graph with permutation-invariant random edges. We present an approximation algorithm for the Balanced Cut problem that finds a balanced cut of cost $O(|E_{\text{random}}|) + n \cdot \text{polylog}(n)$ in this model.

In the most interesting regime, this is a constant factor approximation with respect to the cost of the planted cut.

3.16 Bilu-Linial Stable Instances of Max Cut and Minimum Multiway Cut

Yury Makarychev (TTIC – Chicago, US)

Joint work of Makarychev, Konstantin; Makarychev, Yury; Vijayaraghavan, Aravindan
Main reference K. Makarychev, Y. Makarychev, A. Vijayaraghavan, "Bilu-Linial Stable Instances of Max Cut and Minimum Multiway Cut," in Proc. of the 25th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA'14), pp. 890–906, SIAM, 2014.
URL http://dx.doi.org/10.1137/1.9781611973402.67

We investigate the notion of stability proposed by Bilu and Linial. We obtain an exact polynomial-time algorithm for γ-stable Max Cut instances with $\gamma > c\sqrt{\log n}\log\log n$ for some absolute constant $c > 0$. Our algorithm is robust: it never returns an incorrect answer; if the instance is γ-stable, it finds the maximum cut; otherwise, it either finds the maximum cut or certifies that the instance is not γ-stable. We prove that there is no robust polynomial-time algorithm for γ-stable instances of Max Cut when γ is less than the best approximation factor for Sparsest Cut with non-uniform demands. That suggests that solving γ-stable instances with $\gamma = o(\sqrt{\log n})$ might be difficult or even impossible.

Our algorithm is based on semidefinite programming. We show that the standard SDP relaxation for Max Cut (with ℓ_2^2 triangle inequalities) is integral if $\gamma > D(n)$, where $D(n)$ is the least distortion with which every n point metric space of negative type embeds into ℓ_1. On the negative side, we show that the SDP relaxation is not integral when $\gamma < D(n/2)$. Moreover, there is no tractable convex relaxation for γ-stable instances of Max Cut when γ is less than the best approximation factor for Sparsest Cut.

Our results significantly improve previously known results. The best previously known algorithm for γ-stable instances of Max Cut required that $\gamma > c\sqrt{n}$ (for some $c > 0$) [1]. No hardness results were known for the problem.

Additionally, we present an exact robust polynomial-time algorithm for 4-stable instances of Minimum Multiway Cut. We also study a relaxed notion of weak stability and present algorithms for weakly stable instances of Max Cut and Minimum Multiway Cut.

References

1 Yonatan Bilu, Amit Daniely, Nati Linial, and Michael Saks. On the practically interesting instances of MAXCUT. In *30th Int'l Symp. on Theoretical Aspects of Computer Science*, vol. 20 of *LIPIcs – Leibniz International Proceedings in Informatics*, pp. 526–537, 2013. http://dx.doi.org/10.4230/LIPIcs.STACS.2013.526

3.17 Smoothed Analysis of Local Search

Bodo Manthey (University of Twente, NL)

Local search algorithms are a powerful tool for finding good solutions for optimization problems that often works remarkable well in practice. Yet many local search algorithms show a very bad performance in the worst case.

Smoothed analysis has been applied successfully in order to explain the performance of local search algorithms.

We survey results and open problems of smoothed analysis applied to local search algorithms.

3.18 Analyzing Non-Convex Optimization for Dictionary Learning

Tengyu Ma (Princeton University, US)

Joint work of Arora, Sanjeev; Ge, Rong; Ma, Tengyu; Moitra, Ankur

Dictionary Learning, or as it is often called sparse coding is a basic algorithmic primitive in many machine learning applications, such as image denoising, edge detection, compression and deep learning. Most of the existing algorithms for dictionary learning minimize a non-convex function by heuristics like alternating minimization, gradient descent or their variants. Despite their success in practice, they are not mathematically rigorous because they could potentially converge to local minima.

In this work we show that, alternating minimization and some other variants indeed converge to global minima provably, under a generative model where samples are of the from $A \cdot x$, and A is a incoherent and low spectral norm ground truth dictionary and x is a stochastic sparse vector. Our framework of analysis is potentially useful for analyzing other alternating minimization type algorithms for problems with hidden variables. This is joint work with Sanjeev Arora, Rong Ge and Ankur Moitra.

3.19 Efficient Algorithms Beyond Worst-Case Combinatorial Bounds

Matthias Mnich (Universität Bonn, DE)

A common trait in combinatorics is to find maximum-size or minimum-size substructures of discrete object, and to measure the size of the structure in terms of the size of the object. These combinatorial results often come with polynomial-time algorithms that show how to find the desired structure of optimal size efficiently. For example, in any 3-SAT formula a fraction of 7/8 of its clauses can always be satisfied by some assignment which can be found in polynomial time.

We address the question how to beat the extremal bounds by a small additive constant k, by algorithms whose run-time depends exponentially only on this k. Such kind of problems are called "parameterizations above guarantee". We present some recent results and work in progress in this area.

3.20 Smoothed Analysis of Multi-Objective Optimization

Heiko Röglin (Universität Bonn, DE)

Joint work of Röglin, Heiko; Brunsch, Tobias
Main reference T. Brunsch, H. Röglin, "Improved Smoothed Analysis of Multiobjective Optimization," in Proc. of the 44th Annual ACM Symp./ on Theory of Computing (STOC'12), pp. 407–426, ACM, 2012.
URL http://dx.doi.org/10.1145/2213977.2214016

For most multi-objective optimization problems, the number of Pareto-optimal solutions is usually small in experiments even though in the worst case instances with an exponential number of Pareto-optimal solutions exist. In order to explain this discrepancy, the number of Pareto-optimal solutions has been studied in the model of smoothed analysis in a series of papers over the last decade. In this talk we survey some recent upper and lower bounds on the smoothed number of Pareto-optimal solutions.

3.21 Decompositions of Triangle-Dense Graphs

Tim Roughgarden (Stanford University, US)

Joint work of Gupta, Rishi; Roughgarden, Tim; Seshadhri, C.
Main reference R. Gupta, T. Roughgarden, C. Seshadhri, "Decompositions of Triangle-Dense Graphs," to appear in Proc. of ITCS 2015; pre-print available from author's webpage.
URL http://theory.stanford.edu/~tim/papers/triangle.pdf

High triangle density – the graph property stating that a constant fraction of two-hop paths belong to a triangle – is a common signature of social networks. This paper studies triangle-dense graphs from a structural perspective. We prove constructively that significant portions of a triangle-dense graph are contained in a disjoint union of dense, radius 2 subgraphs. This result quantifies the extent to which triangle-dense graphs resemble unions of cliques. We also show that our algorithm recovers planted clusterings in approximation-stable k-median instances.

3.22 Utilitarian View of Rankings

Or Sheffet (Harvard University, US)

Joint work of Boutillier, Craig; Caragiannis, Ioannis; Haber, Simi; Lu, Tyler; Procaccia, Ariel; Sheffet, Or
Main reference C. Boutilier, I. Caragiannis, S. Haber, T. Lu, A. D. Procaccia, O. Sheffet, "Optimal social choice functions: a utilitarian view," in Proc. of the 13th ACM Conf. on Electronic Commerce (EC'12), pp. 197–214, ACM, 2012; pre-print available from author's webpage.
URL http://dx.doi.org/10.1145/2229012.2229030
URL http://www.cs.cmu.edu/~arielpro/papers/optvoting.full.pdf

In the extensively studied problem of Social Choice (or rank aggregation) n individuals are picking together one alternative out of m possible alternatives. Each individual has a preference among all m alternatives (a total ranking), and the social choice function takes as input these n rankings and outputs a single chosen alternative, called the winner. In this talk we do not focus on the age-old question of truthful rank aggregation, but rather attempt to define the high-level idea that the winner is supposed to be the most favorable alternative for the entire population.

Inspired but newly formulated ideas as to the role of clustering [1], we view social choice as a proxy for maximizing social welfare. Our premise is that agents have (possibly implicit or latent) utility functions, and the goal of a social choice function is to maximize the social welfare – i. e., (possibly weighted) sum of agent utilities – of the selected alternative. We study the model both under worst-case and non-worst case assumptions. We will also discuss current, open ended, work as to maximizing utility of a matching / stable matching.

References

1 Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1068–1077. SIAM, Philadelphia, PA, 2009.

3.23 Probabilistic Lipschitzness: A Measure of Data Niceness

Ruth Urner (Carnegie Mellon University – Pittsburgh, US)

Main reference R. Urner, "Learning with non-Standard Supervision," Ph. D. Thesis, University of Waterloo, 2013.
URL https://uwspace.uwaterloo.ca/handle/10012/7925

Machine learning theory mostly analyses the learning performance of learning in the worst case over all data-generating distributions. However, this framework is often too pessimistic. In particular, for learning settings that employ unlabeled data, such as semi-supervised and active learning, provable savings in label complexity are impossible without assumptions of niceness on the data generating distribution. We propose Probabilistic Lipschitzness (PL), to model marginal-label relatedness of a distribution. This notion is particularly useful for modeling niceness of distributions with deterministic labeling functions. We present convergence rates for Nearest Neighbor learning under PL. We further summarize reductions in labeled sample complexity for learning with unlabeled data (semi-supervised and active learning) under PL.

References

1 Ruth Urner. *Learning with non-Standard Supervision*. Ph. D. Thesis, University of Waterloo, 2013.

2 Ruth Urner and Shai Ben-David and Shai Shalev-Shwartz. *Unlabeled data can Speed up Prediction Time*. Proceedings of the 28th International Conference on Machine Learning (ICML), 2011.

3 Ruth Urner, Sharon Wulff and Shai Ben-David. *PLAL: Cluster-based active learning*. Proceedings of the 26th Conference on Learning Theory (COLT), 2013.

Participants

- Spyros Angelopoulos
CNRS & UPMC, FR
- Sanjeev Arora
Princeton University, US
- Maria-Florina Balcan
Carnegie Mellon University, US
- Shai Ben-David
University of Waterloo, CA
- Markus Bläser
Universität des Saarlandes, DE
- Avrim Blum
Carnegie Mellon University, US
- Karl Bringmann
ETH Zürich, CH
- Kamiel Cornelissen
University of Twente, NL
- Radu Curticapean
Universität des Saarlandes, DE
- Sanjoy Dasgupta
University of California –
San Diego, US
- Matthias Englert
University of Warwick, GB
- Michael Etscheid
Universität Bonn, DE
- Uriel Feige
Weizmann Institute, IL
- Tobias Friedrich
Universität Jena, DE
- Anupam Gupta
Carnegie Mellon University, US
- Marvin Künnemann
MPI für Informatik –
Saarbrücken, DE
- Tengyu Ma
Princeton University, US
- Konstantin Makarychev
Microsoft Res. – Redmond, US
- Yury Makarychev
TTIC – Chicago, US
- Bodo Manthey
University of Twente, NL
- Matthias Mnich
Universität Bonn, DE
- Heiko Röglin
Universität Bonn, DE
- Tim Roughgarden
Stanford University, US
- Or Sheffet
Harvard University, US
- Christian Sohler
TU Dortmund, DE
- Ruth Urner
Carnegie Mellon University, US
- Sergei Vassilvitskii
Google Inc. –
Mountain View, US

Report from Dagstuhl Seminar 14381

Neural-Symbolic Learning and Reasoning

Edited by

Artur d'Avila Garcez[1], Marco Gori[2], Pascal Hitzler[3], and Luís C. Lamb[4]

1 **City University London, GB,** aag@soi.city.ac.uk
2 **University of Siena, IT,** marcoxgori@gmail.com
3 **Wright State University – Dayton, US,** pascal.hitzler@wright.edu
4 **Federal University of Rio Grande do Sul – Porto Alegre, BR,** LuisLamb@acm.org

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 14381 "Neural-Symbolic Learning and Reasoning", which was held from September 14th to 19th, 2014. This seminar brought together specialist in machine learning, knowledge representation and reasoning, computer vision and image understanding, natural language processing, and cognitive science. The aim of the seminar was to explore the interface among several fields that contribute to the effective integration of cognitive abilities such as learning, reasoning, vision and language understanding in intelligent and cognitive computational systems. The seminar consisted of contributed and invited talks, breakout and joint group discussion sessions.

Seminar September 14–19, 2014 – http://www.dagstuhl.de/14381

1998 ACM Subject Classification I.2 Artificial Intelligence, I.2.4 Knowledge Representation Formalisms and Methods, I.2.6 Learning, I.2.10 Vision and Scene Understanding, I.2.11 Distributed Artificial Intelligence

Keywords and phrases Neural-symbolic computation, deep learning, image understanding, lifelong machine learning, natural language understanding, ontology learning

Digital Object Identifier 10.4230/DagRep.4.9.50

1 Executive Summary

Artur S. d'Avila Garcez
Marco Gori
Pascal Hitzler
Luís C. Lamb

Neural-symbolic computation aims at building rich computational models and systems through the integration of connectionist learning and sound symbolic reasoning [1, 2]. Over the last three decades, neural networks were shown effective in the implementation of robust large-scale experimental learning applications. Logic-based, symbolic knowledge representation and reasoning have always been at the core of Artificial Intelligence (AI) research. More recently, the use of deep learning algorithms have led to notably efficient applications, with performance comparable to those of humans, in particular in computer image and vision understanding and natural language processing tasks [3, 4, 5]. Further, advances in fMRI allow scientists to grasp a better understanding of neural functions, leading to realistic neural-computational models. Therefore, the gathering of researchers from several

Neural-Symbolic Learning and Reasoning, *Dagstuhl Reports*, Vol. 4, Issue 9, pp. 50–84
Editors: Artur S. d'Avila Garcez, Marco Gori, Pascal Hitzler, and Luís C. Lamb
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

communities seems fitting at this stage of the research in neural computation and machine learning, cognitive science, applied logic, and visual information processing. The seminar was an appropriate meeting for the discussion of relevant issues concerning the development of rich intelligent systems and models, which can, for instance integrate learning and reasoning or learning and vision. In addition to foundational methods, algorithms and methodologies for neural-symbolic integration, the seminar also showcase a number of applications of neural-symbolic computation.

The meeting also marked the 10th anniversary of the workshop series on neural-symbolic learning and reasoning (NeSy), held yearly since 2005 at IJCAI, AAAI or ECAI. The NeSy workshop typically took a day only at these major conferences, and it became then clear that given that the AI, cognitive science, machine learning, and applied logic communities share many common goals and aspirations it was necessary to provide an appropriately longer meeting, spanning over a week. The desire of many at NeSy to go deeper into the understanding of the main positions and issues, and to collaborate in a truly multidisciplinary way, using several applications (e. g. natural language processing, ontology reasoning, computer image and vision understanding, multimodal learning, knowledge representation and reasoning) towards achieving specific objectives, has prompted us to put together this Dagstuhl seminar marking the 10th anniversary of the workshop.

Further, neural-symbolic computation brings together an integrated methodological perspective, as it draws from both neuroscience and cognitive systems. In summary, neural-symbolic computation is a promising approach, both from a methodological and computational perspective to answer positively to the need for effective knowledge representation, reasoning and learning systems. The representational generality of neural-symbolic integration (the ability to represent, learn and reason about several symbolic systems) and its learning robustness provides interesting opportunities leading to adequate forms of knowledge representation, be they purely symbolic, or hybrid combinations involving probabilistic or numerical representations.

The seminar tackled diverse applications, in computer vision and image understanding, natural language processing, semantic web and big data. Novel approaches needed to tackle such problems, such as lifelong machine learning [6], connectionist applied logics [1, 2], deep learning [4], relational learning [7] and cognitive computation techniques have also been extensively analyzed during the seminar. The abstracts, discussions and open problems listed below briefly summarize a week of intense scientific debate, which illustrate the profitable atmosphere provided by the Dagstuhl scenery. Finally, a forthcoming article describing relevant challenges and open problems will be published at the Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches at the AAAI Spring Symposium Series, to be held at Stanford in March 2015 [8]. This article also adds relevant content and a view of the area, illustrating its richness which may indeed lead to rich cognitive models integrating learning and reasoning effectively, as foreseen by Valiant [9].

Finally, we see neural-symbolic computation as a research area which reaches out to distinct communities: computer science, neuroscience, and cognitive science. By seeking to achieve the fusion of competing views it can benefit from interdisciplinary results. This contributes to novel ideas and collaboration, opening interesting research avenues which involve knowledge representation and reasoning, hybrid combinations of probabilistic and symbolic representations, and several topics in machine learning which can lead to both the construction of sound intelligent systems and to the understanding and modelling of cognitive and brain processes.

References

1 Artur S. d'Avila Garcez, Luis C. Lamb, and Dov M. Gabbay, Neural-Symbolic Cognitive Reasoning. Cognitive Technologies, Springer, 2009.

2 Barbara Hammer, Pascal Hitzler (Eds.): Perspectives of Neural-Symbolic Integration. Studies in Computational Intelligence 77, Springer 2007.

3 D. C. Ciresan, U. Meier, J. Schmidhuber. Multi-column Deep Neural Networks for Image Classification. IEEE Conf. on Computer Vision and Pattern Recognition CVPR 2012.

4 G. E. Hinton, S. Osindero, and Y. Teh, A fast learning algorithm for deep belief nets. Neural Computation, 18, 1527–1554, 2006..

5 Abdel-rahman Mohamed, George Dahl & Geoffrey Hinton. Acoustic Modeling Using Deep Belief Networks. IEEE Transactions on Audio, Speech, and Language Processing. 20(1):14–22, 2012.

6 D. Silver, Q. Yang, and L. Li, Lifelong machine learning systems: Beyond learning algorithms. Proceedings of the AAAI Spring Symposium on Lifelong Machine Learning, Stanford University, AAAI, March, 2013, pp. 49–55.

7 Stephen Muggleton, Luc De Raedt, David Poole, Ivan Bratko, Peter A. Flach, Katsumi Inoue, Ashwin Srinivasan: ILP turns 20 – Biography and future challenges. Machine Learning, 86(1):3–23, 2012.

8 Artur d'Avila Garcez, Tarek R. Besold, Luc de Raedt, Peter Foeldiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kuehnberger, Luis C. Lamb, Risto Miikkulainen, Daniel L. Silver. Neural-Symbolic Learning and Reasoning: Contributions and Challenges. Proceedings of the AAAI Spring Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches, Stanford, March 2015.

9 L. G. Valiant, Knowledge Infusion: In Pursuit of Robustness in Artificial Intelligence. FSTTCS, pp. 415–422, 2008.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

3.1 Symbolic neural networks for cognitive capacities

Tsvi Achler (IBM Almaden Center, US)

Main reference T. Achler, "Symbolic neural networks for cognitive capacities," Biologically Inspired Cognitive Architectures, 9:71–81, 2014.
URL http://dx.doi.org/10.1016/j.bica.2014.07.001

Pattern recognition (identifying patterns from the environment using those stored in memory) and recall (describing or predicting the inputs associated with a stored pattern that can be recognized) are essential for neural-symbolic processing. Without them the brain cannot interact with the world e. g.: understand the environment, logic, and reason. Neural networks are efficient, biologically plausible algorithms that can perform large scale recognition. However, most neural network models of recognition perform recognition but not recall. It remains difficult to connect models of recognition with models of logic and emulate fundamental brain functions, because of the symbolic recall limitation. Before discussing symbolic networks further, one of the important realizations from the Dagstuhl seminar is that folks that focus on neural networks have a different definition of symbolic (and sub-symbolic) than folks that focus on logic. This matter was not fully solved. Subsequently I carefully define symbolic and note that in some literatures this term may be used differently. Here symbolic (call it "functional" symbolic?) is defined by the relation between input features and outputs (e. g. zebra has 4 legs). I assume that weights of neurons responsible for zebra demark mark this in connection weights that do not change. Let me clarify. There are two types of neural networks in the literature defined by how neurons learn for recognition processing: localist and globalist. In localist methods only neurons related to the information adjust their weights based learning on rules quantified within the neuron. Simple Hebbian learning is an example of this rule. Globalist methods in contrasts may require all neurons (including those that are not directly responsible) to change their weights to learn a new relation. PDP and feedforward models are examples of global learning. My symbolic definition is localist because I assumed the zebra neuron is independent of other neurons in that it does not change if another neuron is added with another symbolic relation (e. g. there exists another neuron representing another animal that has 0,4,6,8 or however many legs). Using this definition a neural network that is symbolic neural network cannot be globalist. A symbolic network also requires the ability to recall: to be able to derive from the symbol (e. g. zebra) what are the characteristic components (e. g. 4 legs, stripes etc). Thus the label (e. g. zebra) behaves as a symbol that encapsulates the components that are associated with it (legs, stripes, tail, hooves etc). Globalist networks cannot recall and subsequently in some literatures are called sub-symbolic (e. g. [2, 3]). Fortunately localist networks involve symmetrical top-down connections (from label to components) and the best example of such networks are auto-associative networks (e. g. Restricted Boltzmann Machines for Deep Learning). However auto-associative networks have self-excitatory symmetrical connections (positive feedback). A property of self-excitatory feedback is that iterative activation of even small values will lead to the maximal values regardless whether non-binary values are used. This degrades performance. I introduce a different localist model from auto-associative networks that uses are self- inhibitory symmetrical connections (negative feedback). The proposed model can converge to non-binary real-valued activations and is sensitive to real-valued weights. Moreover the network can be shown mathematically to obtain analogous solutions as standard feedforward (globalist) neural networks. Thus we have a model that can be as powerful as popular globalist neural networks, but is localist and symbolic. It can perform recall: retrieve the components involved in recognizing the label [1]. I hope to see more focus on these type of approaches within the neural symbolic community.

References

1 Achler T. Symbolic Neural Networks for Cognitive Capacities, Special issue of the Journal on Biologically Inspired Cognitive Architectures, 9, 71–81, 2014.

2 Fodor J.A., Pylyshyn Z.W. Connectionism and Cognitive Architecture: A Critical Analysis, Cognition 28, 3–71, 1988.

3 Sun R. Duality of the Mind: A Bottom-up Approach Toward Cognition. Mahwah, NJ: Lawrence Erlbaum Associates, 2002.

3.2 On extracting Rules for: enriching ontological knowledge bases, complementing heterogeneous sources of information, empowering the reasoning process

Claudia d'Amato (University of Bari, IT)

Joint work of d'Amato, Claudia; Staab, Steffen; Bryl, Volha; Serafini, Luciano

The Linked Open Data (LOD) cloud, which represents a significant example of Bid Data, could be seen as a huge portion of assertional knowledge whose intentional part is formally defined by existing OWL ontologies freely available on the Web. LOD constitutes a tremendous source of knowledge, that as such needs effective and efficient methods for its management. Data mining techniques could play a key role with this respect. The focus of the talk is on the discovery and extraction of knowledge patterns that are hidden in the (often noisy and inherently incomplete) data. Hidden knowledge patterns are extracted in the form of (relational) association rules by exploiting the evidence coming from the ontological knowledge bases [1] and/or from heterogeneous sources of information (i. e. an ontology and a relational databases referring to the same domain) [2] as well as by exploiting reasoning capabilities. While using methods at the state of the art, that as such necessarily need a further and deeper investigation for really scaling on very large data sets, the main focus will be on the potential that the extracted rules may have for: enriching existing ontological knowledge bases, for complementing heterogeneous sources of information, and for empowering the deductive reasoning process.

Particularly, the talk is organized in two parts. In the first one, the focus is on extracting hidden knowledge patterns from purely ontological knowledge bases. In the second one, the focus is on extracting hidden knowledge patterns from heterogeneous source of information.

The key observation motivating the first part of the talk is given by the fact that ontological knowledge bases are often not complete in the sense that missing concept and role assertions, with respect to the reference domain, can be found, as well as missing disjointness axioms and/or relationships. In order to cope with this problem, a method for discovering DL-Safe [4, 5] *Relational Association rules*, represented with SWRL [3] language, is presented [1]. This method is intended to discover all possible hidden knowledge patters that may be used for: a) (semi-)automatizing the completion of the assertional knowledge (given the pattern in the left hand side of a discovered rule, a new concept/role assertion may be induced by the right hand side of the rule); b) straightforwardly extending and enriching the expressive power of existing ontologies with formal rules, while ensuring and maintaining the decidability of the reasoning operators (because DL-Safe SWRL rules are extracted [3, 5]); c) suggesting knew knowledge axioms (induced by the discovered association rules). Inspired to [11, 12], the proposed method implements a level-wise generate-and-test approach that, starting with an initial general pattern, i. e. a concept name (jointly with a variable name) or a role name (jointly with variable names) proceeds, at each level, with generating a number of specializations by the use of suitable operators defined for the purpose. Each specialized pattern is then evaluated, on the ground of formally defined conditions, for possible pruning. This process is iterated until a predefined stopping criterion met. Besides of developing a scalable algorithm, the experimental evaluation of the developed method represents one of the most challenging problem since it requires the availability of gold standards (currently not available) with respect to which assessing the validity of the induced new knowledge. A possible solution is presented in [6].

As regards the second part of the talk, the motivating observation is given by the fact that even if available domain ontologies are increasing over the time, there is still a huge amount of data stored and managed with RDBMS and referring to the same domain. The key idea is that this complementarity could be exploited for discovering knowledge patterns that are not formalized within the ontology (or the RDBMS) but that are learnable from the data. For the purpose, a framework for extracting hidden knowledge patterns across ontologies and relational DBMS, called *Semantically Enriched Association Rules*, is illustrated [2, 13]. It is grounded on building an integrated view of (a part of) the RDBM and the ontology in a tabular representation which allows the exploitation of well know state of the art algorithms, such as the APRIORI algorithm [14], for extracting Association Rules. The extracted patterns can be used for enriching the available knowledge (in both format) and for refining existing ontologies. Additionally, the extracted semantically enriched association rules can be exploited when performing deductive reasoning on an ontological knowledge bases. Specifically, a modified Tableaux algorithm, that we call *Data Driven Tableaux algorithm* is introduced [15, 13]. It is intended as a method for performing automated reasoning on grounded knowledge bases (i. e. knowledge bases linked to RDBMS data) which combines logical reasoning and statistical inference (coming from the discovered semantically enriched association rules) thus making sense of the heterogeneous data sources. The goals of the Data Driven Tableaux algorithm are twofold. On one hand it aims at reducing the computational effort for finding a model for a given (satisfiable) concept. On the other hand it aims at suppling the "most plausible model", that is the one that best fits the available data, for a given concept description. The key point of the algorithm is a defined heuristic, exploiting the semantically enriched association rules, to be used when random choices (e. g. when processing a concepts disjunction) occur. The proposed framework has to be intended as the backbone of a mixed models representation and reasoning.

The exploitation of association rules is not new in the Semantic Web context. In [6], a framework for discovering association rules for predicting new role assertions from an RDF data source is proposed, but no reasoning capabilities and TBox information are exploited for the purpose. Additionally, the extracted patterns are not integrated in the considered source of knowledge. Heterogeneous sources of information have been considered in [7, 8], where frequent patterns are discovered, respectively in the form of DATALOG clauses, from an $\mathcal{AL}$-Log knowledge base at different granularity level, and in the form of conjunctive queries, given a specified objective. Additional usages of association rules have been proposed in [9], where association rules are learnt from RDF data for inducing a schema ontology, but without exploiting any reasoning capabilities and in [10] where association rules are exploited for performing RDF data compression.

References

1 C. d'Amato and S. Staab. Predicting Assertions in Ontological Knowledge Bases by Discovering Relational Association Rules. Technical Report. http://www.di.uniba.it/~cdamato/TechAssoRules.pdf (2013).

2 C. d'Amato and Volha Bryl and Luciano Serafini. Semantic Knowledge Discovery from Heterogeneous Data Sources. In Proc. of the Knowledge Engineering and Knowledge Management – 18th International Conference (EKAW'12), vol. 7603, pp. 26–31, Springer, LNCS (2012).

3 I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean. Swrl: A semantic web rule language combining owl and ruleml http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/, 2004.

4 B. Motik, U. Sattler, and R. Studer. Query answering for owl-dl with rules. In *Proceedings of the International Semantic Web Conceference*, volume 3298 of *LNCS*, pages 549–563. Springer, 2004.

5 B. Motik, U. Sattler, and R. Studer. Query answering for owl-dl with rules. *Web Semantics*, 3(1):41–60, 2005.

6 L. Galárraga, C. Teflioudi, F. Suchanek, and Katja Hose. Amie: Association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 20th International World Wide Web Conference (WWW 2013).* ACM, 2013.

7 F. A. Lisi. Al-quin: An onto-relational learning system for semantic web mining. *International Journal of Semantic Web and Information Systems*, 2011.

8 J. Józefowska, A. Lawrynowicz, and T. Lukaszewski. The role of semantics in mining frequent patterns from knowledge bases in description logics with rules. *Theory and Practice of Logic Programming*, 10(3):251–289, 2010.

9 J. Völker and M. Niepert. Statistical schema induction. In G. Antoniou et al., editors, *The Semantic Web: Research and Applications – 8th Extended Semantic Web Conference, (ESWC 2011), Proc., Part I*, volume 6643 of *LNCS*, pages 124–138. Springer, 2011.

10 A. K. Joshi and P. Hitzler and G. Dong Logical Linked Data Compression In *The Semantic Web: Research and Applications – 10th Extended Semantic Web Conference, (ESWC 2013), Proceedings*, volume 7882 of *LNCS*, pages 170–184. Springer, 2013.

11 L. Dehaspeand and H. Toironen. Discovery of frequent datalog patterns. *Data Mining and Knowledge Discovery*, 3(1):7–36, 1999.

12 B. Goethals and J. Van den Bussche. Relational association rules: Getting warmer. In *Proceedings of the International Workshop on Pattern Detection and Discovery*, volume 2447 of *LNCS*, pages 125–139. Springer, 2002.

13 C. d'Amato and Volha Bryl and Luciano Serafini. Semantic Knowledge Discovery and Data-Driven Logical Reasoning from Heterogeneous Data Sources. In F. Bobillo et al.(Eds.), ISWC International Workshops, URSW 2011-2013, Revised Selected Papers, vol. 8816 Springer, LNCS/LNAI (2014).

14 R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proc. of the 1993 ACM SIGMOD Int. Conf. on Management of Data*, pages 207–216. ACM Press, 1993.

15 C. d'Amato and Volha Bryl and Luciano Serafini. Data-Driven Logical Reasoning. In Proc. of the 8th Int. Workshop on Uncertainty Reasoning for the Semantic Web (URSW'12), vol. 900, CEUR (2012).

3.3 Neural-Symbolic Computing, Deep Logic Networks and Applications

Artur d'Avila Garcez (City University London, GB)

Joint work of cf. references
Main reference A. S. d'Avila Garcez, L. C. Lamb, D. M. Gabbay, "Neural-Symbolic Cognitive Reasoning," Springer, ISBN 978-3-642-09229-9, 2009.
URL http://www.springer.com/computer/ai/book/978-3-540-73245-7

In this talk I reviewed the work carried out with many collaborators over the past 15 years in the area of neural-symbolic computing, starting with the CILP system for integrating logic programming and recurrent neural networks trained with backpropagation [1]. CILP networks take advantage of background knowledge during learning, which can improve training performance as shown in power systems and bioinformatics applications [2]. Knowledge extraction allows CILP networks to be described in symbolic form for the sake of transfer learning and explanation [3]. Extensions of CILP, including the use of feedback, network ensembles and nested networks, allows the representation and learning of various forms of nonclassical reasoning, including modal, temporal and epistemic reasoning [4, 5], as well as abduction [6]. This has led to a full solution in connectionist form of the so-called muddy children puzzle in logic [7]. Fibring of CILP networks offers further expressive power by combining networks of networks for simultaneous learning and reasoning [8]. Applications have included training and assessment in simulators, normative reasoning and rule learning, integration of run-time verification and adaptation, action learning and description in videos [9, 10, 13]. Current developments and efforts have been focused on: fast relational learning using neural networks (the CILP++ system) [11] and effective knowledge extraction from large networks, including deep networks and the use of knowledge extraction for transfer learning [12]. Future applications include the analysis of complex networks, social robotics and health informatics, and multimodal learning and reasoning combining video and audio data with metadata.

References

1 A. S. d'Avila Garcez, K. Broda and D. M. Gabbay, *Neural-Symbolic Learning Systems: Foundations and Applications*, Springer, 2002.

2 A. S. d'Avila Garcez and G. Zaverucha. The Connectionist Inductive Learning and Logic Programming System. Applied Intelligence 11(1):59–77, 1999.

3 A. S. d'Avila Garcez, K. Broda and D. M. Gabbay. Symbolic Knowledge Extraction from Trained Neural Networks: A Sound Approach. Artificial Intelligence, 125(1–2):153–205, January 2001.

4 A. S. d'Avila Garcez and L. C. Lamb. A Connectionist Computational Model for Epistemic and Temporal Reasoning. Neural Computation, 18(7):1711–1738, July 2006.

5 A. S. d'Avila Garcez, L. C. Lamb and D. M. Gabbay. Connectionist Modal Logic: Representing Modalities in Neural Networks. Theoretical Computer Science, 371(1–2):34–53, February 2007.

6 A. S. d'Avila Garcez, D.M. Gabbay, O. Ray and J. Woods. Abductive Reasoning in Neural-Symbolic Learning Systems. Topoi: An International Review of Philosophy, 26:37–49, March 2007.

7 A. S. d'Avila Garcez, L. C. Lamb and D. M. Gabbay. *Neural-Symbolic Cognitive Reasoning*, Springer, 2009.

8 A. S. d'Avila Garcez and D. M. Gabbay. Fibring Neural Networks. In Proc. 19th National Conference on Artificial Intelligence AAAI 2004. San Jose, California, USA, AAAI Press, July 2004.

9 R. V. Borges, A. S. d'Avila Garcez and L. C. Lamb. Learning and Representing Temporal Knowledge in Recurrent Networks. IEEE Transactions on Neural Networks, 22(12):2409–2421, December 2011.

10 L. de Penning, A. S. d'Avila Garcez, L. C. Lamb and J. J. Meyer. A Neural-Symbolic Cognitive Agent for Online Learning and Reasoning. In Proc. IJCAI'11, Barcelona, Spain, July 2011.

11 M. Franca, G. Zaverucha and A. S. d'Avila Garcez. Fast Relational Learning using Bottom Clause Propositionalization with Artificial Neural Networks, Machine Learning, 94(1):81–104, 2014.

12 S. Tran and A. S. d'Avila Garcez. Logic Extraction from Deep Belief Networks. In Proc. ICML Workshop on Representation Learning, Edinburgh, Scotland, July 2012.

13 G. Boella, S. Colombo-Tosatto, A. S. d'Avila Garcez, V. Genovese, A. Perotti and L. van der Torre. Learning and Reasoning about Norms using Neural-Symbolic Systems. In Proc. 11th International Conference on Autonomous Agents and Multiagent Systems, AAMAS'12, Valencia, Spain, June 2012.

3.4 Dreaming and Consciousness in Deep Neural-Symbolic Cognitive Agents

Leo de Penning (TNO Behaviour and Societal Sciences – Soesterberg, NL)

Joint work of de Penning, Leo; Garcez, Artur d'Avilla; Meyer, John-Jules
Main reference L. de Penning, A. d'Avilla Garcez, J.-J. C. Meyer, "Dreaming Machines: On multimodal fusion and information retrieval using Neural-Symbolic Cognitive Agents," in Proc. of the 2013 Imperial College Computing Student Workshop (ICCSW'13), OASIcs, Vol. 35, pp. 89–94, 2013.
URL http://dx.doi.org/10.4230/OASIcs.ICCSW.2013.89

Deep Boltzmann Machines (DBM) have been used as a computational cognitive model in various AI-related research and applications, notably in computational vision and multimodal fusion. Being regarded as a biological plausible model of the human brain, the DBM is also becoming a popular instrument to investigate various cortical processes in neuroscience. In this paper, we describe how a multimodal DBM is implemented as part of a Neural-Symbolic Cognitive Agent (NSCA) for real-time multimodal fusion and inference of streaming audio and video data. We describe how this agent can be used to simulate certain neurological mechanisms related to hallucinations and dreaming and how these mechanisms are beneficial to the integrity of the DBM. Also we will explain how the NSCA is used to extract multimodal information from the DBM and provide a compact and practical iconographic temporal logic formula for complex relations between visual and auditory patterns. Finally we will discuss the implications of the work in relation to Machine Consciousness.

3.5 Progress in Probabilistic Logic Programming

Luc De Raedt (KU Leuven, BE)

Probabilistic logic programs combine the power of a programming language with a possible world semantics, typically based on Sato's distribution semantics and they have been studied for over twenty years. In this talk, I introduced the concepts underlying probabilistic programming, their semantics, different inference and learning mechanisms. I then reported on recent progress within this paradigm. This was concerned with an extension towards dealing with continuous distributions as well as coping with dynamics. This is the framework of distributional clauses that has been applied to several applications in robotics, for tracking relational worlds in which objects or their properties are occluded in real time. Finally, some remaining open challenges were discussed.

See also the websites http://dtai.cs.kuleuven.be/problog/ and http://dtai.cs.kuleuven.be/ml/systems/DC/ for more details and an interactive tutorial on ProbLog.

References

1 Nitti, D., De Laet, T., & De Raedt, L. (2013). A particle filter for hybrid relational domains. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2013.

2 De Raedt, L., & Kimmig, A. Probabilistic Programming Concepts, arXiv:1312.4328, 2013.

3.6 Semantic and Fuzzy Modelling and Recognition of Human Activities in Smart Spaces. A case study on Ambient Assisted Living

Natalia Díaz-Rodríguez (Turku Centre for Computer Science, FI)

Joint work of Díaz-Rodríguez, Natalia; León Cadahía, Olmo; Cuéllar, M. P.; Lilius, Johan; Delgado Calvo-Flores, Miguel

Main reference N. Díaz-Rodríguez, O. León Cadahía, M. P. Cuéllar, J. Lilius, M. Delgado Calvo-Flores, "Handling real-world context-awareness, uncertainty and vagueness in real-time human activity tracking and recognition with a fuzzy ontology-based hybrid method," MDPI Physical Sensors, 14(10):18131–18171, 2014.

URL http://dx.doi.org/10.3390/s141018131

Human activity recognition in everyday environments is a critical task in Ambient Intelligence applications to achieve proper Ambient Assisted Living. Key challenges still remain to be tackled to achieve robust methods. Our hybrid system allows to model and recognize a set of complex scenarios where vagueness and uncertainty is inherent to the human nature of the users that perform it. We provide context meaning to perform sub- activity tracking and recognition from depth video data. To achieve a more loosely coupled model that lets flexibility to be part of the recognition process, we validate the advantages of a hybrid data-driven and knowledge-driven system with a challenging public dataset and achieve an accuracy of 90.1% and 91.1% respectively for low and high-level activities. The handling of uncertain, incomplete and vague data (i. e., missing sensor readings or execution variations) is tackled for first time with a public depth-video dataset taking into account the semantics of activities, sub-activities and real-time object interaction. This entails an improvement over both entirely data-driven approaches and merely ontology- based approaches.

3.7 Making the latent category structure of fMRI data explicit with Formal Concept Analysis

Dominik Endres (Universität Marburg, DE)

Joint work of Endres, Dominik; Adam, Ruth; Giese, Martin A.; Noppeney, Uta
Main reference D. Endres, R. Adam, M. A. Giese, U. Noppeney, "Understanding the Semantic Structure of Human fMRI Brain Recordings With Formal Concept Analysis," in Proc. of the 10th Int'l Conf. on Formal Concept Analysis (ICFCA'12), LNCS, Vol. 7278, pp. 96-111, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-29892-9_13

Understanding how semantic information is represented in the brain has been an important research focus of Neuroscience in the past few years. The work I presented in this talk is aimed at extracting concepts and their relationships from brain activity, and to correlate these concept with behavioral measures. We showed previously (Endres et al 2010) that Formal Concept Analysis (FCA) can reveal interpretable semantic information (e. g. specialization hierarchies, or feature-based representations) from electrophysiological data. Unlike other analysis methods (e. g. hierarchical clustering), FCA does not impose inappropriate structure on the data. FCA is a mathematical formulation of the explicit coding hypothesis (Foldiak, 2009) Furthermore we (Endres et al 2012) investigated whether similar findings can be obtained from fMRI BOLD responses recorded from human subjects. While the BOLD response provides only an indirect measure of neural activity on a much coarser spatio-temporal scale than electrophysiological recordings, it has the advantage that it can be recorded from humans, which can be questioned about their perceptions during the experiment. Furthermore, the BOLD signal can be recorded from the whole brain simultaneously. In our experiment, a single human subject was scanned while viewing 72 grayscale pictures of animate and inanimate objects in a target detection task. These pictures comprise the formal objects for FCA. We computed formal attributes by learning a hierarchical Bayesian classifier, which maps BOLD responses onto binary features, and these features onto object labels. The connectivity matrix between the binary features and the object labels can then serve as the formal context. In a high-level visual cortical area (IT), we found a clear dissociation between animate and inanimate objects with the inanimate category subdivided between animals and plants when we increased the number of attributes extracted from the fMRI signal. The inanimate objects were hierarchically organized into furniture and other common items, including vehicles. We also used FCA to display organizational differences between high-level and low-level visual processing areas. For a quantitative validation of that observation, we show that the attribute structure computed from the IT fMRI signal is highly predictive of subjective similarity ratings, but we found no such relationship to responses from early visual cortex. Collaborators: Peter Foldiak, Uta Priss, Ruth Adam, Uta Noppeney

References

1 D. Endres, P. Földiak and U. Priss (2010) An Application of Formal Concept Analysis to Semantic Neural Decoding, Annals of Mathematics and Artificial Intelligence, 57(3–4), pp. 233–248.

2 D. Endres, R. Adam, M. A. Giese and U. Noppeney (2012) Understanding the Semantic Structure of Human fMRI Brain Recordings With Formal Concept Analysis., In ICFCA 2012, 10th Inte'l Con. on Formal Concept Analysis, LNCS 7278, Springer, 96–111.

3 P. Földiak (2009) Neural Coding: Non-Local but Explicit and Conceptual, Current Biology, 19(19):R904–R906, 2009.

4 B. Ganter and R. Wille (1999). Formal concept analysis: Mathematical foundations, Springer.

3.8 Symbolic Data Mining Methods Applied to Human Sleep Records

Jacqueline Fairley (Emory University – Atlanta, US)

Joint work of Fairley, Jacqueline; Georgoulas, George; Karvelis, Petros; Stylios, Chrysostomos; Rye, David; Bliwise, Donald
Main reference J. A. Fairley, G. Georgoulas, P. Karvelis, C. D. Stylios, D. B. Rye, D. L. Bliwise, "Symbolic Representation of Human Electromyograms for Automated Detection of Phasic Activity during Sleep," 2014.
URL https://www.academia.edu/7659893/SYMBOLIC_REPRESENTATION_OF_HUMAN_ELECTROMYOGRAMS_FOR_AUTOMATED_DETECTION_OF_PHASIC_ACTIVITY_DURING_SLEEP

Background: Phasic electromyographic (EMG)/muscle activity in human overnight polysomnograms (PSGs) represent a potential indicator/quantitative metric for identifying various neurodegenerative disorder populations and age-matched controls [1].

Unfortunately, visual labeling of phasic EMG activity is time consuming making this method unscalable for clinical implementation. Therefore, we propose computerized labeling of EMG activity in a detection scheme utilizing k-Nearest Neighbor classification and Symbolic Aggregate approXimation (SAX), a novel algorithm from the field of time series data mining that transforms a time series, such as EMG, into a string of arbitrary symbols [2]. A primary advantage of SAX analysis includes access to robust symbolic based data mining algorithms viable for scalable computing.

Methods: Six male subjects (S001:S006) polysomnograms (PSGs)/sleep data sets were visually scored, using one second epochs, for phasic and non-phasic left and right leg EMG activity (sampling rate 200Hz), by the same trained visual scorer. Phasic muscle activity epochs were characterized by amplitudes visually exceeding four times the surrounding background activity and having time durations between 100 to 500 msec. SAX was applied to all EMG records using a one second non-overlapping moving window, four symbol alphabet, and $\frac{1}{2}$ sec frames, followed by translation of the SAX string into an intelligent icon, a color mapped image representing the frequency of each word in the SAX string. Results: SAX based classification scheme results, using 10-fold cross validation and k- Nearest Neighbor Classification (best of k:1:1:7; minimum value:increment value:maximum value), were compared to visual labeling [3]. Detection of non-phasic EMG activity exceeded 90% for all six subjects: S001 (98.4), S002 (97.8), S003 (98.1), S004 (93.6), S005 (95.2), and S006 (95.8). Phasic EMG activity detection surpassed 80% for three subjects: S001 (90.5), S004 (81.8), and S006 (87.1). However, phasic EMG activity detection decreased in performance for S002 (61.0), S003 (53.6) and S005 (68.0).

Conclusions: Detection rates for half of the subjects indicate feasibility of replacing tedious expert visual scoring with the proposed computational scheme. However, this scheme lacks robustness across all subjects, and requires refinement of SAX alphabet size and frame length along with comparison with other classification algorithms such as Support Vector Machines and Random Forest. Most importantly, efficient fine-tuning of this computational scheme promises to hasten computerized EMG activity scoring for neurodegenerative disorder tracking in clinical settings.

References

1 Bliwise, D. L., He, L., Ansari, F. P., and Rye, D. B. (2006) Quantification of electromyographic activity during sleep: a phasic electromyographic metric. Journal of Clinical Neurophysiology, Vol. 23, pp. 59–67.

2 Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003) A Symbolic Representation of Time Series, with Implications for Streaming Algorithms, In Proc. of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego, CA.

3 Fairley, J. A., Georgoulas, G., Karvelis, P., Stylios, C. D., Rye, D. B., Bliwise, D. L. (2014). Symbolic Representation of Human Electromyograms for Automated Detection of Phasic Activity during Sleep, Statistical Signal Processing Workshop (SSP) Proceedings, 2014 IEEE, pp. 185–188, Gold Coast, Australia.

3.9 Affordances, Actionability, and Simulation

Jerry A. Feldman (ICSI – Berkeley, US)

The notion of affordances depends crucially on the actions available to an agent in context. When we add the expected utility of these actions in context, the result has been called actionability. There is increasing evidence that AI and Cognitive Science would benefit from shifting from a focus on abstract "truth" to treating actionability as the core issue for agents. Actionability also somewhat changes the traditional concerns of affordances to suggest a greater emphasis on active perception. An agent should also simulate (compute) the likely consequences of actions by itself or other agents. In a social situation, communication and language are important affordances.

3.10 Simulation Semantics and the Rebirth of NLU

Jerry A. Feldman (ICSI – Berkeley, US)

Natural Language Understanding (NLU) was one of the main original goals of artificial intelligence and cognitive science. This has proven to be extremely challenging and was nearly abandoned for decades. We describe an implemented system that supports full NLU for tasks of moderate complexity. The natural language interface is based on Embodied Construction Grammar and simulation semantics. The system described here supports dialog with an agent controlling a simulated robot, but is flexible with respect to both input language and output task.

3.11 The Neural Binding Problem(s)

Jerry A. Feldman (ICSI – Berkeley, US)

As with many other "problems" in vision and cognitive science, "the binding problem" has been used to label a wide range of tasks of radically different behavioral and computational structure. These include a "hard" version that is currently intractable, a feature-binding variant that is productive routine science and a variable-binding case that is unsolved, but should be solvable. The talk will cover all these and some related problems that seem intractably hard as well as some that are unsolved, but are being approached with current and planned experiments.

References

1 Jerome A. Feldman. *The Neural Binding Problem(s)*. Cognitive Neurodynamics, Vol. 7, Issue 1, pp. 1–11, February 2013

3.12 Fast Relational Learning using Neural Nets

Manoel Franca (City University London, GB)

Joint work of Franca, Manoel; Garcez, Artur; Zaverucha, Gerson
URL http://www.dagstuhl.de/mat/Files/14/14381/14381.FrancaManoel.ExtAbstract.pdf

Relational learning can be described as the task of learning first-order logic rules from examples. It has enabled a number of new machine learning applications, e. g. graph mining and link analysis. We introduce a fast method and system for relational learning, called CILP++, which handles first-order logic knowledge and have been on several ILP datasets, comparing results with Aleph. The results show that CILP++ can achieve accuracy comparable to Aleph, while being generally faster. Several alternative approaches, both for BCP propositionalization and for CILP++ learning, are also investigated.

3.13 Evolutionary and Swarm Computing for the Semantic Web

Christophe D. M. Gueret (DANS – Den Hague, NL)

Joint work of Gueret, Christophe D. M.; Schlobach, S. ; Dentler, K. ; Schut, M. ; Eiben, G.
Main reference C. Gueret, S. Schlobach, K. Dentler, M. Schut, G. Eiben, "Evolutionary and Swarm Computing for the Semantic Web," IEEE Computational Intelligence Magazine, 7(2):16–31, 2012.
URL http://dx.doi.org/10.1109/MCI.2012.2188583

The Semantic Web has become a dynamic and enormous network of typed links between data sets stored on different machines. These data sets are machine readable and unambiguously interpretable, thanks to their underlying standard representation languages. The expressiveness and flexibility of the publication model of Linked Data has led to its widespread adoption and an ever increasing publication of semantically rich data on the Web. This success however has started to create serious problems as the scale and complexity of information outgrows the current methods in use, which are mostly based on database technology, expressive knowledge representation formalism and high-performance computing. We argue that methods from computational intelligence can play an important role in solving these problems. In this paper we introduce and systemically discuss the typical application problems on the Semantic Web and argue that the existing approaches to address their underlying reasoning tasks consistently fail because of the increasing size, dynamicity and complexity of the data. For each of these primitive reasoning tasks we will discuss possible problem solving methods grounded in Evolutionary and Swarm computing, with short descriptions of existing approaches. Finally, we will discuss two case studies in which we successfully applied soft computing methods to two of the main reasoning tasks; an evolutionary approach to querying, and a swarm algorithm for entailment.

3.14 Computer Science for Development

Christophe D. M. Gueret (DANS – Den Hague, NL)

Joint work of Gueret, Christophe D. M.; Schlobach, S. ; de Boer V.
Main reference World Wide Semantic Web community, "Let's build a world wide Semantic Web!," 2011.
URL http://worldwidesemanticweb.org

Data sharing usually focuses on centralized and very powerful solutions centred around Web hosted servers and (mobile) clients accessing it. As a direct consequence, the usage of Linked Data technology depends on the availability of a Web infrastructure compassing data-centres, high speed reliable Internet connection and modern client devices. If any of this is missing, our community is not able, yet, to provide any Linked Data enabled data management solution. Still, the digital divide that is currently widely recognized separates the world into those who have access to Web-based platforms and those who don't. When designing Linked Data platforms we tend to forget those 4 Billion persons who don't have access to Internet but would benefit from being able to share structured data. We should keep everyone in mind when we design Linked Data platforms and aim at helping to reduce this digital divide. We believe that achieving this goal implies working on three aspects (Infrastructure, Interfaces and Relevancy) around open data.

This problem the Semantic Web community faces doing knowledge representation in developing countries is only one facet of Computer Science. Many other aspects of it are also concerned. For instance, Human-Computer Interaction (HCI) need to account for users that don't read or write or don't speak any "common" language, Engineering need to be performed on smaller scale devices with sparse networkings and Information retrieval need to be done with a focus on locally relevant information. These many aspects of Computer Sciences affected by the specific challenges posed by using ICT in the developing world call for a global study over CS4D where researchers would join in ensuring the technology they work on is inclusive and usable by everyone world wide.

3.15 Combining Learning and Reasoning for Big Data

Pascal Hitzler (Wright State University – Dayton, US)

Reasoning and learning are natural allies. The former provides deductive expert system-like capabilities for dealing with interpretation of data, while the latter focuses on finding patterns in data. This perspective suggests a rather obvious workflow in which inductive and statistical methods analyze data, resulting in metadata which describes higher-level conceptualizations (metadata) of the data, which in turn enables the use of the data and metadata in deduction-based systems. However, this apparently obvious pipeline is broken since the current state of the art leaves gaps which need to be bridged by new innovations.

In this presentation, we discuss some of the recent work which addresses these gaps, with the goal of stimulating further research on the interplay between learning and reasoning.

References

1 P. Hitzler and F. van Harmelen. A reasonable semantic web. *Semantic Web*, 1(1–2):39–44, 2010.

2 P. Jain, P. Hitzler, P. Z. Yeh, K. Verma, and A. P. Sheth. Linked Data is Merely More Data. In D. Brickley, V. K. Chaudhri, H. Halpin, and D. McGuinness, editors, *Linked Data Meets Artificial Intelligence*, pages 82–86. AAAI Press, Menlo Park, CA, 2010.

3 P. Hitzler, M. Krötzsch, and S. Rudolph. *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC, 2010.

4 P. Hitzler and K. Janowicz. Linked Data, Big Data, and the 4th Paradigm. *Semantic Web*, 4(3):233–235, 2013.

5 K. Janowicz and P. Hitzler. The Digital Earth as knowledge engine. *Semantic Web*, 3(3):213–221, 2012.

6 P. Jain, P. Hitzler, A. P. Sheth, K. Verma, and P. Z. Yeh. Ontology alignment for linked open data. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *The Semantic Web – ISWC 2010 – 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7–11, 2010, Revised Selected Papers, Part I*, volume 6496 of *Lecture Notes in Computer Science*, pages 402–417. Springer, 2010.

7 A. Joshi, P. Jain, P. Hitzler, P. Yeh, K. Verma, A. Sheth, and M. Damova. Alignment-based querying of Linked Open Data. In R. Meersman, H. Panetto, T. Dillon, S. Rinderle-Ma, P. Dadam, X. Zhou, S. Pearson, A. Ferscha, S. Bergamaschi, and I. F. Cruz, editors, *On the Move to Meaningful Internet Systems: OTM 2012, Confederated International Conferences: CoopIS, DOA-SVI, and ODBASE 2012, Rome, Italy, September 10–14, 2012, Proceedings, Part II*, volume 7566 of *Lecture Notes in Computer Science*, pages 807–824. Springer, Heidelberg, 2012.

3.16 From Human Reasoning Episodes to Connectionist Models

Steffen Hölldobler (TU Dresden, DE)

URL http://www.wv.inf.tu-dresden.de/~sh/

I present a new approach to model human reasoning based on reasoning towards an appropriate logical form, weak completion semantics, three-valued Lukasiewicz logic, and an appropriate semantic operator. The approach admits least models and, hence, reasoning is performed with respect to least models. After adding abduction the approach can adequately handle human reasoning episodes like the suppression and the selection task. Moreover, it can be mapped into a connectionist model using the core method.

3.17 On Concept Learning as Constructive Reasoning

Francesca Alessandra Lisi (University of Bari, IT)

Main reference F. A. Lisi, "A Declarative Modeling Language for Concept Learning in Description Logics," in Proc. of the 22nd Int'l Conf. on Inductive Logic Programming (ILP'12), LNCS, Vol. 7842, pp. 151–165, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-38812-5_11

In this talk I provided a novel perspective on Concept Learning, which relies on recent results in the fields of Machine Learning (ML)/Data Mining (DM) and Knowledge Representation (KR), notably De Raedt *et al.*'s work on declarative modeling of ML/DM problems [2] and Colucci *et al.*'s work on non-standard reasoning in the KR framework of Description Logics (DLs) [1]. In particular, I provided a formal characterization of Concept Learning which arises from the observation that the inductive inference deals with finding – or constructing – a concept. More precisely, non-standard reasoning services which support the inductive inference can be modeled as constructive reasoning tasks where the solution construction may be subject to optimality criteria. Under this assumption, I defined a declarative language – based on second-order DLs – for modeling different variants of the Concept Learning problem (namely, Concept Induction, Concept Refinement and Concept Formation) [3]. The language abstracts from the specific algorithms used to solve the Concept Learning problem in hand. However, as future work, I emphasized the need for an efficient and/or effective solver to make the proposed language more attractive from a practical point of view.

References

1 Colucci, S., Di Noia, T., Di Sciascio, E., Donini, F. M., Ragone, A.: A unified framework for non-standard reasoning services in description logics. In: Coelho, H., Studer, R., Wooldridge, M. (eds.) ECAI 2010 – 19th European Conference on Artificial Intelligence, Lisbon, Portugal, August 16–20, 2010, Proceedings. Frontiers in Artificial Intelligence and Applications, vol. 215, pp. 479–484. IOS Press (2010)

2 De Raedt, L., Guns, T., Nijssen, S.: Constraint programming for data mining and machine learning. In: Fox, M., Poole, D. (eds.) Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11–15, 2010. AAAI Press (2010)

3 Lisi, F. A.: A declarative modeling language for concept learning in description logics. In: Riguzzi, F., Zelezny, F. (eds.) Inductive Logic Programming, 22nd International Conference, ILP 2012, Dubrovnik, Croatia, September 17–19, 2012, Revised Selected Papers. Lecture Notes in Computer Science, vol. 7842. Springer Berlin Heidelberg (2013)

3.18 Interactive Intelligent Systems: Scaling Learning with the Expert in the Loop

Dragos Margineantu (Boeing Research & Technology, US)

URL http://www.dmargineantu.net/

Research in intelligent neural and symbolic systems has made significant advances with respect to the accuracy of predictions, detections, classifications. However in order to deploy these algorithms and tools, to execute or assist the execution of real world tasks, in most of the cases, these methods require the assistance of an AI expert. A suite of practical tasks can be addressed optimally at this point in time by a team that combines the expertise of the user with the strength of automated intelligent systems. Can we develop (or adapt our) existing algorithms for such tasks? We believe so! By formulating our research questions to capture the expert-intelligent system goals. This presentation will show how we formulated the research questions and adapted techniques such as inverse reinforcement learning (IRL) or active learning for assisting experts in tasks such as detecting abnormal agent behavior, scene analysis, and estimating intent. We will also outline some open research questions for usable expert-interactive learning.

3.19 Concepts, Goals and Communication

Vivien Mast (Universität Bremen, DE)

Joint work of Mast, Vivien;Couto Vale, Daniel; Falomir, Zoe; Elahi, Fazleh Mohammad
Main reference V. Mast, V., D. Couto Vale, Z. Falomir, F. M. Elahi, "Referential Grounding for Situated Human-Robot Communication," in Proc. of the 18th Workshop on the Semantics and Pragmatics of Dialogue (SemDial'14), pp. 223–225, 2014.
URL http://www.macs.hw.ac.uk/InteractionLab/Semdial/semdial14.pdf

Much work in computational linguistics and cognitive science implicitly rests on the idea, dating back to Plato and Aristotle, that there are rational categories which are sets of entities in the real world, defined by necessary and sufficient properties, and that the power of linguistic expressions and mental concepts stems from their correspondence to such rational categories. I will discuss some limitations of a rational notion of concepts and meaning in the domain of reference, and argue that human concepts should be viewed from the perspective of actionability, as suggested by Jerry Feldman at this seminar. In particular, I will argue that concept assignment depends on context and the goals of the conceptualizing agent.

In the standard paradigm of REG (Krahmer & van Deemter, 2012), objects are represented by attribute-value pairs. The task of REG is defined as finding, for a given target object, a distinguishing description – a set of attribute-value pairs whose conjunction is true of the target but not of any of the other objects in the domain. However, research on collaborative reference has shown that reference ultimately does not rely on truth, but on common ground and efficient grounding mechanisms (Clark & Bangerter, 2004). I will argue that meta-knowledge about the potential of conceptual mismatch and miscommunication guide concept assignment in reference, and I will present the Probabilistic Reference And GRounding mechanism PRAGR for generating and interpreting referring expressions (Mast et al., 2014; Mast & Wolter, 2013). PRAGR is geared towards maximizing mutual understanding by flexibly assigning linguistic concepts to objects, depending on context.

References

1 Clark, H. H. & Bangerter, A. *Changing Ideas about Reference*. In: Sperber, Dan and Noveck, Ira A. (Ed.), Experimental Pragmatics. Hampshire, NY: Palgrave Macmillan, 2004.

2 Krahmer, E. & van Deemter, K. *Computational Generation of Referring Expressions: A Survey*. Computational Linguistics, 38(1), pp. 173–218, 2012.

3 Mast, V., Couto Vale, D., Falomir, Z. & Elahi, F. M. *Referential Grounding for Situated Human-Robot Communication.*. Proceedings of DialWatt/SemDial. 2014.

4 Mast, V. & Wolter, D. *A Probabilistic Framework for Object Descriptions in Indoor Route Instructions.*. In Tenbrink, T. and Stell, J. and Galton, A. and Wood, Z. (Ed.), Spatial Information Theory (Vol. 8116, pp. 185–204). Springer International Publishing, 2013

3.20 Grounding Meaning in Perceptual Representations

Risto Miikkulainen (University of Texas – Austin, US)

Joint work of Miikkulainen, Risto; Aguilar, Mario; Aguirre-Celis, Nora; Binder, Jeffrey R.; Connolly, Patrick; Fernandino, Leo; Morales, Isaiah; Williams, Paul;

How word meaning may be grounded in perceptual experience is a fundamental problem in neural-symbolic learning. I will describe an artificial neural network model that shows how this process may take place through learned associations between visual scenes and linguistic phrases. I will then describe ongoing work on identifying such associations from fMRI images of sentence comprehension.

3.21 Mining Graphs from Event Logs

Andrey Mokhov (Newcastle University, GB)

Joint work of Mokhov, Andrey; Carmona, Josep
Main reference A. Mokhov, J. Carmona, "Process Mining using Parameterised Graphs," Technical memo, Newcastle University, August 2014.
URL http://async.org.uk/tech-memos/NCL-EEE-MICRO-MEMO-2014-009.pdf

We introduce a mathematical model for compact representation of large families of (related) graphs [1], detecting patterns in graphs, and using such compact representations for process mining [2]. By process mining we mean understanding or explanation of behaviour of complex systems by observing events occurring in them. These events come in the form of event logs that record event types, time stamps and other associated metadata. The task of process mining is to extract useful knowledge from such logs, for example, to explain, predict or diagnose complex systems. We present graph-theoretic methods that extract information about concurrency and causality from such logs, and then attempt to represent the result in the most compact/simple form hopefully amenable to human understanding [3].

References

1 Andrey Mokhov, Alex Yakovlev. *Conditional partial order graphs: Model, synthesis, and application*, IEEE Transactions on Computers 59 (11), 1480–1493, 2010.

2 Wil van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, Springer, 2011.

3 Andrey Mokhov, Victor Khomenko. *Algebra of Parameterised Graphs*, ACM Transactions on Embedded Computing, 2014.

3.22 Learning Compositional Robot Activities from Examples

Bernd Neumann (Universität Hamburg, DE)

Joint work of Neumann, Bernd; Hotz, Lothar; Rost, Pascal; Lehmann, Jos
Main reference B. Neumann, L. Hotz, P. Rost, J. Lehmann, "A Robot Waiter Learning from Experiences," in Proc. of the 10th Int'l Conf. on Machine Learning and Data Mining in Pattern Recognition (MLDM'14), LNCS, Vol. 8556, pp. 286–299, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-08979-9_22

In the KR framework of the EU project RACE, as in many other systems, robot activities are described by compositional hierarchies connecting activity concepts at higher abstraction levels with components at lower levels, down to action primitives of the robot platform with quantitative parameters, and down to percepts at neural level. One way for a service robot to increase its competence is to learn new activities based on known subactivities and coarse instructions. Given an initial repertoire of basic operations, such a process can establish compositional structures at increasingly high levels of abstraction and complexity. In this talk I describe recent advances in learning compositional structures using a Description Logic (DL) extended by semantic attachments as formal knowledge representation framework. A learning curriculum, based on positive examples, is presented where the robot has to determine autonomously which spatiotemporal conditions must be satisfied for a newly learnt activity. It is shown that the robot can construct conceptual descriptions from the examples in such a way that the intended target description is approached with monotonously increasing generality. The generalization process is realized by aligning concept graphs obtained from DL representations and merging corresponding nodes by a Good Common Subsumer (GCS). It is shown that this process can also be used for adapting an existing concept to a new situation. Examples are presented for a service robot learning waiter activities in a restaurant domain.

3.23 Neural-Symbolic Runtime Verification

Alan Perotti (University of Turin, IT)

Joint work of Perotti, Alan; d'Avila Garcez, Artur; Boella, Guido
Main reference A. Perotti, A. S. d'Avila Garcez, G. Boella, "Neural Networks for Runtime Verification," in Proc. of the 2014 Int'l Joint Conf. on Neural Networks (IJCNN'14), pp. 2637–2644, IEEE, 2014.
URL http://dx.doi.org/10.1109/IJCNN.2014.6889961

I introduced RuleRunner, a novel Runtime Verification system for monitoring LTL properties over finite traces. By exploiting results from the Neural-Symbolic Integration area, a RuleRunner monitor can be encoded in a recurrent neural network. The results show that neural networks can perform real-time runtime verification and techniques of parallel computing can be applied to improve the performance in terms of scalability. Furthermore, our framework allows for property adaptation by using a standard neural network learning algorithm.

3.24 Symbolic Computation, Binding and Constraint Learning in Bolzmann Machines

Gadi Pinkas (Center for Academic Studies, Or-Yehudah and Bar-Ilan University, IL)

Joint work of Pinkas, G, Cohen S., Lima P.
Main reference G. Pinkas, P. Lima, S. Cohen, "Representing, binding, retrieving and unifying relational knowledge using pools of neural binders," Journal of Biologically Inspired Cognitive Architectures, 6(October 2013):87–95, 2013.
URL http://dx.doi.org/10.1016/j.bica.2013.07.005

For a long time, connectionist architectures have been criticized for having propositional fixation, lack of compositionality and, in general, for their weakness in representing sophisticated symbolic information, learning it and processing it. This work offers an approach that allows full integration of symbolic AI with the connectionist paradigm. We show how to encode, learn and process relational knowledge using attractor based artificial neural networks, such as Boltzmann Machines. The neural architecture uses a working memory (WM), consisting of pools of "binders", and a long-term synaptic-memory (LTM) that can store a large relational knowledge-base (KB). A compact variable binding mechanism is proposed which dynamically allocates ensembles of neurons when a query is clamped; retrieving KB items till a solution emerges in the WM. A general form of the Hebbian learning rule is shown that learns from constraint violations. The learning rule is applied to High-Order Boltzmann machines (with sigma-pi connections) and is shown to learn networks with attractors (energy minima) representing correct symbolic inferences. We illustrate the mechanism using predicate logic inference problems and planning in block-world.

The mechanism uses a biologically inspired cognitive architecture, which is based on relatively compact Working Memory and larger synaptic Long-Term-Memory which stores knowledge that constrains the neural activation of the WM and forms attractors in its dynamics. In this architecture, knowledge items are retrieved from LTM into the WM only upon need, and, graph-like structures, that represent solution inferences, emerge at thermal equilibrium as an activation pattern of the neural units. Our architecture is based on the fact that Boltzmann Machines may be viewed as performing constraint satisfaction, where, at equilibrium, fixed-points maximally satisfy a set of weighted constraints. We show how to encode and bind arbitrary complex graphs as neural activation in WM and how a supervised learner may use miscalculations to adjust synapses so that constraints are better enforced, in order to correctly retrieve and process such complex structures. The architecture allows learning representations as expressive as First-Order-Logic (with bounded proof length), has no central control and is inherently robust to unit failures. The mechanism is goal directed in the sense, that the query may drive the processing, as well as the current activation pattern in the WM. It is universal and has a simple underlying computational principle. As such, it may be further adapted for applications that combine the advantages of both connectionist and traditional symbolic AI and may be used in modeling aspects of human' cognition.

3.25 Learning Action-oriented Symbols: Abstractions over Decision Processes

Subramanian Ramamoorthy (University of Edinburgh, GB)

Joint work of Ramamoorthy, Subramanian; Mahmud, M. M. Hassan; Hawasly, Majd; Rosman, Benjamin; Pokorny, Florian T.

A key question at the interface between sub-symbolic and symbolic learning and reasoning is that of how symbols can be acquired from experience, and grounded. Can such symbols be action-oriented in that they consistently abstract the underlying process?

I discuss two approaches we have recently developed for categorising policies obtained through processes such as reinforcement learning or motion planning in robots. The goal of categorisation is to arrive at a set of action- relevant symbols that better enable reasoning about changes associated with dynamic environments; taking a transfer/lifelong learning perspective.

The first approach is to cluster decision processes in terms of similarities in the effects of the actions. We define a novel distance and a clustering algorithm that yields a smaller set of decision processes that make continual transfer algorithms more effective.

The second approach draws on new mathematical tools from computational topology to abstract a set of trajectories associated with motion plans, yielding entirely qualitative descriptions of the underlying domain – which can again be used to separate quantitative detail from other global structural aspects of the tasks. I end by asking how these principles can be incorporated with a variety of models being studies by the NeSy community, including in particular deep networks.

References

1 M. M. H. Mahmud, M. Hawasly, B. Rosman, S. Ramamoorthy, Clustering Markov Decision Processes for continual transfer, arXiv:1311.3959 [cs.AI].

2 F. T. Pokorny, M. Hawasly, S. Ramamoorthy, Multiscale topological trajectory classification with persistent homology, In Proc. *Robotics: Science and Systems* (R:SS), 2014.

3.26 Mixing Low-Level and Semantic Features for Image Interpretation: A framework and a simple case study

Luciano Serafini (Fondazione Bruno Kessler and University of Trento, IT)

Joint work of Donadello, Ivan; Serafini, Luciano
Main reference I. Donadello, L. Serafini, "Mixing low-level and semantic features for image interpretation," in Proc. of the 1st Int'l Workshop on Computer Vision and Ontologies, to appear.

In recent years internet has seen a terrific increase of digital images. Thus the need of searching for images on the basis of human understandable descriptions, as in the case of textual documents, is emerging. For this reason, sites as YouTube, Facebook, Flickr, Grooveshark allow the tagging of the media and support search by keywords and by examples. Tagging activity is very stressful and often is not well done by users. For this reason automatic methods able to automatically generate a description of the image content, as in textual documents, become a real necessity. There are many approaches to image understanding which try to generate a high level description of an image by analysing low-level information (or features), such as colours, texture and contours, thus providing such a high level description in terms of semantic concepts, or high-level information. This would allow a person to search, for instance, for an image containing "a man is riding an horse". The difficulty to find the correspondence between the low-level features and the human concepts is the main problem in content-based image retrieval. It is the so-called *semantic gap* [2]. It's widely recognised that, to understand the content of an image, contextual information (aka background knowledge) is necessary [3]. Background knowledge, relevant to the context of an image, can be expressed in terms of logical languages in an ontology [4]. In image interpretation ontologies can be used for two main purposes. First, ontologies allow the expression of a set of constraints on the possible interpretations which can be constructed by considering only low-level features of an image. The satisfaction of such constraints can be checked via logical reasoning. Second, the terminology introduced in the ontology can be used as formal language to describe the content of the images. This will enable semantic image retrieval using queries expressed in the language introduced by the ontology. The background knowledge formalizes the semantics of the human understandable concepts and will provide the set of types of objects that can be found in a picture (e. g., horse, human, etc.) and the set of relations that can exist between depicted objects (e. g., rides is a relation between a human and an animal, part-of is a general relation between physical objects, etc.). Furthermore, the background knowledge provides constraints on types of objects and relations, e. g. a vehicle has at least two wheels or horses are animals that can be ridden by men. The advantage of having the tags as concepts coming from a background knowledge allows to reason over the image. For example the tag "horse" enables to infer the presence of an animal.

In the present work we adopt the natural idea that, already introduced for instance in [5, 6, 7] where an interpretation of a picture, in the context of an ontology, is a (partial) model of the ontology itself that expresses the state of affairs of the world in the precise moment in which the picture has been taken. We propose to formalize the notion of image interpretation, w.r.t. an ontology, as *a segmented image, where each segment is aligned with an object of a partial model of the reference ontology.* To cope with the fact that a picture reports only partial information on the state of affairs we use the notion of partial model of a logical theory [8]; to cope with the possibility of having multiple alternative interpretations of a picture we introduce the notion of *most plausible interpretation* an image, which is the interpretation that maximises some scoring function.

In order to have a preliminary evaluation of our idea, we implemented this framework, for a specific and limited case. We developed a fully unsupervised method to generate image interpretations able to infer the presence of complex objects from the parts present in the picture, thus inferring the relative "part-whole" structure. The method jointly exploits the constraints on the part-whole relation given by the ontology, and the low-level features of the objects available in the image. From a preliminary evaluation the presented approach shows promising results.

References

1 Donadello, I., Serafini, L.: Mixing low-level and semantic features for image interpretation. In: 1st International Workshop on Computer Vision and Ontologies at ECCV 2014, Zurich, Switzerland, Sept, 2014. (2014) To appear.

2 Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recognition **40**(1) (January 2007) 262–282

3 Oliva, A., Torralba, A.: The role of context in object recognition. Trends in cognitive sciences **11**(12) (2007) 520–527

4 Bannour, H., Hudelot, C.: Towards ontologies for image interpretation and annotation. In Martinez, J.M., ed.: 9th International Workshop on Content-Based Multimedia Indexing, CBMI 2011, Madrid, Spain, June 13–15, 2011, IEEE (2011) 211–216

5 Reiter, R., Mackworth, A.K.: A logical framework for depiction and image interpretation. Artificial Intelligence **41**(2) (1989) 125–155

6 Neumann, B., Möller, R.: On scene interpretation with description logics. Image and Vision Computing **26**(1) (2008) 82–101 Cognitive Vision-Special Issue.

7 Straccia, U.: Reasoning within fuzzy description logics. J. Artif. Intell. Res. (JAIR) **14** (2001) 137–166

8 Staruch, B., Staruch, B.: First order theories for partial models. Studia Logica **80**(1) (2005) 105–120

3.27 Lifelong Machine Learning and Reasoning

Daniel L. Silver (Acadia University – Wolfville, CA)

URL http://plato.acadiau.ca/courses/comp/dsilver

Lifelong Machine Learning (LML) considers systems that learn many tasks over a lifetime, accurately and efficiently retaining and consolidating the knowledge they have learned and using that knowledge to more quickly and accurately learn new tasks [2, 1]. Since 1999, I have investigated aspects of LML for Learning to Classify (L2C) problem domains. In [3] I provide an overview of prior work in LML, present a framework for LML, and discuss its two essential ingredients – knowledge retention [4] and transfer learning [1]. Transfer learning is about using prior knowledge to more accurately develop models for a new task, from fewer training examples and in shorter periods of time. Knowledge retention is about efficient and effective methods of storing learned models for use in transfer learning and potentially reasoning. The proposed research program extends my prior work on LML to the learning of knowledge for purposes of reasoning. I am motivated by the belief that intelligent agents, like humans, should develop in their abilities as a function of their experience.

My previous research has focused on the theory and application of transfer learning and knowledge consolidation. We have published results on functional and representational knowledge transfer using multiple task learning (MTL), task rehearsal using synthesized training examples, and selective transfer for classification and regression problems [2]. Most significantly, we have developed context-sensitive MTL (csMTL); a transfer learning method that uses an additional context input, rather than an additional output for each new task [Silver09]. This approach overcomes a number of significant problems of standard MTL when applied to a LML

Our research has shown that knowledge of a new task can be integrated, or consolidated, with that of prior tasks in order for a LML solution to overcome the stability-plasticity problem and scale for practical use [4]. The stability-plasticity problem is the loss of prior task knowledge in a neural network when learning the examples of a new task [5]. Our work has demonstrated that MTL and csMTL networks can mitigate this problem by maintaining functional accuracy of prior tasks (stability) through the relearning, or rehearsal, of prior task examples while modifying the representation of the network (plasticity) through the learning new task examples. This can be accomplished using the back-propagation (BP) algorithm under the conditions described in [4, 5]. Recently, we have shown that a mix of proper selection of task rehearsal examples and more advanced methods of regularization can improve consolidation in csMTL networks.

In 2013, Qiang Yang and I encouraged the machine learning community to move beyond learning algorithms to systems that are capable of learning, retaining and using knowledge over a lifetime [3]. ML now has many practical applications of L2C; the next generation of ML needs to consider the acquisition of knowledge in a form that can be used for more general AI, such as Learning to Reason (L2R). We argue that opportunities for advances in AI lie at the locus of machine learning and knowledge representation; specifically, that methods of knowledge consolidation will provide insights into how to best represent knowledge for use in future learning and reasoning.

A survey of ML methods that create knowledge representations that can be used for learning and reasoning revealed three major bodies of work. The first is Neural-Symbolic Integration (NSI) [6, 8]. NSI research considers the benefits of integrating robust neural network learning with expressive symbolic reasoning capabilities. Much of NSI work focuses on the extraction of symbolic rules from trained network weights and the transfer of knowledge from logical expressions to network weights prior to training. Since the early 2000s, members of this community have called for a joint treatment of learning and reasoning [7]. At the IJCAI 2013 NeSy'13 workshop I presented an invited talk on the common ground shared by LML and NSI. I proposed an integrated framework for NSI and LML and discussed how the requirement of reasoning with learned knowledge places an additional constraint on the representational language and search methods used by LML systems. Learning is necessary to acquire knowledge for reasoning, however, reasoning informs us about the best ways to store and access knowledge. Thus, learning and reasoning are complimentary and should be studied together. Recent work at CMU on the NELL system agrees with this combined view [9]. The second major body of work is Learning to Reason (L2R) [10, 11], also referred to as Knowledge Infusion [12, 14, 15]. L2R work is not as abundant as that of NSI; however, it suggests a promising approach to developing is most promising in terms of our proposed research. The L2R framework is concerned with both learning a knowledge representation and with it doing deductive reasoning. The perspective is that an agent only needs to learn the knowledge required to reason in his environment, and to the level of performance demanded by that environment. Unlike prior approaches to engineering common knowledge, such as Cyc [16], L2R takes a probabilistic perspective on learning and reasoning. An L2R agent need not answer all possible knowledge queries, but only those that are relevant to the environment of the agent in a probably approximately correct (PAC) sense; that is, assertions can be learned to a desired level of accuracy with a desired level of confidence [12]. In [10] and [12] both authors show that a L2R framework allows for efficient learning of Boolean logical assertions in the PAC-sense (polynomial in the number of variables and training examples). Further to this, they prove that the knowledge learned can be used to efficiently reason with a similar level of accuracy and confidence. In this way, L2R agents are robust learners, acquiring most accurately the common knowledge that they need to reason in accord with their environment [12]. The authors make the point that traditional artificial intelligence has chosen knowledge representations for their transparency (e. g. preferring CNF over DNF representations) whereas the L2R framework chooses knowledge representations because they are learnable and facilitate reasoning. The third body of work is Deep Learning Architectures (DLA) and includes recent publications on Semi-supervised Learning [17], Co-training [18], Self-taught Learning [24], Representation Learning [20, 21], and Deep Learning [25, 26, 28, 22]. All share a common interest with LML in that they develop knowledge representations of the world from examples that can be used for future learning. This fall we are finalizing a survey of transfer learning and consolidation methods using DLAs.

My future research goals are to (1) develop and test Lifelong Machine Learning and Reasoning (LMLR) systems that can retain learned knowledge in a form that can be used for reasoning as well as future learning; and to (2) study the practical benefits and limitations of a prototype LMLR system applied to real-world problems in data mining and intelligent agents. To advance on the first goal, we will develop a system that can learn a series of logic assertions, such as $A|B \Rightarrow C$ and $C \Rightarrow D$, from examples of those expressions. The resulting knowledge base model can then be used to reason that $A \Rightarrow D$ by testing the model with examples. To advance on the second goal, I will scale the system up such that it can learn to reason from images that encode similar assertions. Such a system could be used by an intelligent agent to provide recommendations on next best action.

This work will create new theory on the learning and representation of knowledge from examples acquired from the learner's environment and methods by which to reason using that learned knowledge. Finding solutions to consolidating new with prior knowledge from examples that contain only part of the input space will be a major challenge. The methods and findings will be of interest to researchers working on machine learning, knowledge representation, reasoning, and applied areas such as data mining and intelligent agents.

References

1 Silver, D. and Mercer, R. 2009. Life-long Learning Through Task Rehearsal and Selective Knowledge Transfer. Theory and Novel Applications of Machine Learning, Editor: Meng Joo Er and Yi Zhou, IN-Tech Education and Pub, pp. 335–356.

2 Silver, D. and Poirier, R. and Currie, D. 2008. Inductive Transfer with Context-Sensitive Neural Networks. Machine Learning – Special Issue on Inductive Transfer, Springer, 73(3):313–336.

3 Silver, D., Yang, Q. and Li, L. Lifelong machine learning systems: Beyond learning algorithms. Proceedings of the AAAI Spring Symposium on Lifelong Machine Learning, Stanford University, AAAI, March, 2013, pp. 49–55.

4 Fowler, B. and Silver, D. 2011, Consolidation using Context-Sensitive Multiple Task Learning, Advances in Artificial Intelligence, 24th Conference of the Canadian Artificial Intelligence Association (AI 2011), St. John's, Nfld, May, 2011, P. Lingras (Ed.), Springer, LNAI 6657, pp. 128–139.

5 Silver, D. The Consolidation of Task Knowledge for Lifelong Machine Learning. Proceedings of the AAAI Spring Symposium on Lifelong Machine Learning, Stanford University, CA, AAAI, March, 2013, pp. 46–48.

6 Artur S. d'Avila Garcez, Luis C. Lamb and Dov M. Gabbay (2009). Neural-Symbolic Cognitive Reasoning. Cognitive Technologies, Springer, 2009.

7 Sebastian Bader, Pascal Hitzler, Steffen Hölldobler (2006). The Integration of Connectionism and First-Order Knowledge Representation and Reasoning as a Challenge for Artificial Intelligence. Information 9 (1).

8 Luis C. Lamb (2008). The Grand Challenges and Myths of Neural- Symbolic Computation. Dagstuhl Seminar on Recurrent Neural Networks, 2008.

9 A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr. and T.M. Mitchell (2010). Toward an Architecture for Never-Ending Language Learning. In Proceedings of the 2010 Conference on Artificial Intelligence (AAAI-2010).

10 R. Khardon and D. Roth (1997), Learning to Reason, Journal of the ACM , vol. 44, 5, pp. 697–725.

11 R. Khardon and D. Roth (1999), Learning to Reason with a Restricted View, Machine Learning , vol. 35, 2, pp. 95–117.

12 Valiant, L. G. (2008). Knowledge Infusion: In Pursuit of Robustness in Artificial Intelligence. In FSTTCS (pp. 415–422).

13 Leslie G. Valiant (2012). A neuroidal architecture for cognitive computation. J. ACM 47(5):854–882.

14 Juba, Brendan (2012). Learning implicitly in reasoning in PAC- Semantics. arXiv preprint arXiv:1209.0056.

15 Juba, Brendan (2013). Implicit learning of common sense for reasoning. IJCAI- 2013.

16 Douglas Lenat and R. V. Guha (1990). Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley.

17 Chapelle, Olivier; Schölkopf, Bernhard; Zien, Alexander (2006). Semi-supervised learning. Cambridge, Mass, MIT Press.

18 Blum, A., Mitchell, T (1998). Combining labeled and unlabeled data with co-training. COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann, pp. 92–100.

19 Bengio, Y. (2011). Deep learning of representations for unsupervised and transfer learning. In JMLR W&CP: Proc. Unsupervised and Transfer Learning.

20 Yoshua Bengio, Aaron Courville and Pascal Vincent (2013). Representation Learning: A Review and New Perspectives in: Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35:8(1798–1828)

21 Yoshua Bengio (2013), Deep learning of representations: looking forward, in: Statistical Language and Speech Processing, pages 1–37, Springer.

22 Adam Coates, Andrej Karpathy, and Andrew Y. Ng (2012). Emergence of Object-Selective Features in Unsupervised Feature Learning. In NIPS 2012.

23 Raina, R., Madhavan, A., and Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In L. Bottou and M. Littman, editors, ICML 2009, pages 873-880, New York, NY, USA.

24 Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. (2007). Self- taught learning: transfer learning from unlabeled data. In ICML-2007.

25 Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18, 1527–1554.

26 Hinton, G. E. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504–507.

27 Hinton, G., Krizhevsky, A., and Wang, S. (2011). Transforming auto- encoders. In ICANN-2011.

28 Le, Q.; Ranzato, M.; Monga, R.; Devin, M.; Chen, K.; Corrado, G.; Dean, J.; and Ng, A. (2012) Building high-level features using large scale unsupervised learning. In Proc. of International Conference in Machine Learning.

3.28 Representation Reuse for Transfer Learning

Son Tran (City University London, GB)

Joint work of Tran, Son; Artur d'Avila Garcez

The recent success of representation learning is built upon the learning of relevant features, in particular from unlabelled data available in different domains. This raises the question of how to transfer and reuse such knowledge effectively so that the learning of a new task can be made easier or be improved. This poses a difficult challenge for the area of transfer learning where there is no label in the source data, and no source data is ever transferred to the target domain. In previous work, the most capable approach has been self- taught learning which, however, relies heavily upon the compatibility across the domains. In this talk, I propose a novel transfer learning framework called Adaptive Transferred-profile Likelihood Learning (aTPL), which performs transformations on the representations to be transferred, so that they become more compatible with the target domain. At the same time, it learns supplementary knowledge about the target domain. Experiments on five datasets demonstrate the effectiveness of the approach in comparison with self- taught learning and other common feature extraction methods. The results also indicate that the new transfer method is less reliant on source and target domain similarity, and show how the proposed form of adaptation can be useful in the case of negative transfer.

3.29 Decoding the Symbols of Life: Learning Cell Types and Properties from RNA Sequencing Data

Joshua Welch (University of North Carolina – Chapel Hill, US)

Recent breakthroughs in biochemical techniques for low-input sequencing have enabled whole-transcriptome quantification (RNA-seq) of single cells. This technology enables molecular biologists to dissect the relationship between gene expression and cellular function at unprecedented resolution. In particular, the cell type composition of tissues is now open to investigation. There is a need for unsupervised learning approaches that can identify cell types and properties from single-cell RNA sequencing data in a purely unbiased manner, rather than relying on previously known cell type markers. This task of identifying cell types and the relationships between them is not unlike recognizing symbols in text or images. An overview of the relevant biological questions, single-cell RNA sequencing technology, and existing approaches to solving this problem are presented. The goal of the talk is to initiate discussion about how neural- symbolic approaches can be used to identify cell types and their properties from single-cell RNA sequencing data.

4 Working Groups

During the workshop several working groups were formed, and lively discussions on relevant research challenges took place. Next, we briefly summarize the results and questions raised during the breakout sessions.

4.1 Consolidation of learned knowledge

This discussion session was related to the concept of Learning to Reason, as investigated by Valiant, Khardon, Roth and many others. Significant advances in AI lie at the locus of machine learning and knowledge representation; specifically, methods of knowledge consolidation will provide insights into how to best represent common knowledge for use in future learning and reasoning. Knowledge consolidation is about efficient and effective methods of sequentially storing knowledge as it is learned. Overcoming the stability-plasticity problem is the main challenge here. Consolidation in neural networks can occur through a slow process of interleaved learning of a new and old task examples within a large multiple task learning (MTL) network. Task rehearsal is one approach to overcoming the stability-plasticity problem of forgetting of previously learned tasks stored in a MTL network by relearning synthesized examples of those tasks while simultaneously learning a new task. However this method faces scaling problems as the number of prior task examples increases. This session discussed new approaches to overcoming the stability plasticity problem so that knowledge consolidation is tractable over long sequences of learning.

4.2 Affordabilities and actionability

Inspired by points raised by Feldman, this session started from the ancient idea that the goal of thought is "truth", which has been productive, but it is also limiting. There are multiple reasons to believe that replacing "truth" with "actionability" will be more fruitful and that this move is necessary for a unified cognitive science. For more on this topic the reader is invited to the work on affordances by Feldman: ftp://ftp.icsi.berkeley.edu/pub/feldman/affordances.jf.pdf

4.3 Closing the gap in the pipeline: how to use learned knowledge for reasoning

Deductive and inductive approaches are natural allies. The former uses high-level conceptualizations to logically reason over data, while the latter focuses on finding higher-level patterns in data. This perspective suggests a rather obvious workflow in which inductive and statistical methods analyze data, resulting in metadata which describes higher level features of the data, which in turn enables the use of the data in intelligent systems. However, this apparently obvious pipeline is broken since the current state of the art leaves gaps which need to be bridged by new innovations. It would be helpful to start establishing the exact nature of these gaps, and to brainstorm about ways how to address these. Advances on this topic should provide added value for large-scale data management and analysis.

4.4 What is symbolic, what is sub-symbolic?

An old debate took place: what is the meaning of the terms symbolic and sub-symbolic in neural computation? Several questions were raised and analyzed. Certain neural networks are symbolic while others are not. What are factors that determine this? How can recognition be performed with symbolic networks? How can recall necessary for reasoning be performed with non-symbolic networks? How can both recognition and recall be achieved with the same networks?

4.5 How far can nature inspire us?

Among all the nature-inspired computation techniques, neural networks are about the only ones to have made their way into knowledge representation and reasoning so far. What about swarm computing or evolutionary computing? Could they also have a role and a use for learning and reasoning problems? The conclusion is that this is a discussion we can have looking at existing prototypes and ongoing research, including recent progress in the area of autonomous agents and multi-agent systems.

4.6 Demos & Implementation Fest

Finally, a hands on session took place. A lively display of neural-symbolic tools was presented by a number of the seminar's participants. The participants had the opportunity to showcase their NeSy related software and get others to try, evaluate and discuss their work. Future extensions and integrations of the showcased work were proposed. Most participants had the opportunity to experiment with existing tools and prototypes that use state-of-the-art neural-symbolic computation techniques for image, audio, video and multimodal learning and reasoning.

5 Open Problems

After each discussion session, challenges and open problems were identified. It is clear that a number of research avenues lay ahead of the communities that participated in the seminar. The list below reflects, in part, the interdisciplinary nature of the research presented and the open problems identified at the seminar, leading to interesting future developments and applications. A companion paper [9] complements the list below and also identifies several opportunities and challenges for the neural-symbolic community.

- Over the last decades, most of the work has been focused on propositional approaches, which was seen as *propositional fixation* by McCarthy [1]. However, novel approaches have significantly contributed to the representation of other logical systems in neural networks, leading to successful application in temporal specification and synchronization [2, 3], distributed knowledge representation [4, 5] and even fragments of first-order logic inference [6]. In order to make progress in this open problem, perhaps one should consider logics of intermediate expressiveness such as description logics of the Horn family [7]. There remains a number of open issues in knowledge representation and reasoning in neural networks, in particular with regard to learning. The integration of neural-symbolic systems and inductive logic programming [8] may also lead to relevant developments. The companion paper [9] also identifies challenges in this area.
- Recently, it has been shown that neural networks are able to learn sequences of actions, a point raised by Icard during the discussions. Thus, it may well be possible that a "mental simulation" of some concrete, temporally extended activity can be effected by connectionist models. Theories of action, based on propositional dynamic logic can thus be useful. Feldman in [10] has argued that if the brain is not a network of neurons that represent things, but a network of neurons that do things, action models would probably be central in this endeavour.
- With respect to how the brain actually represents knowledge, perhaps one can draw inspiration from advances in fMRI. The work of Endres and Foldiak [11] may lead to a biologically sound model of the brain's semantic structures. It can also contribute to the construction of new learning algorithms, by contributing to identifying the functioning of the brain's learning mechanisms.

- There is much work to be done with respect to learning to reason (L2R) in neural networks [12, 13]. A question raised by Silver is how a L2R agent can develop a complete knowledge-base over time when examples of the logical expressions arrive with values for only part of the input space. Perhaps a Lifelong Machine Learning (LML) approach is needed. Such an approach can integrate, or consolidate, the knowledge of individual examples over many learning episodes [14]. Consolidation of learned knowledge is a necessary requirement as it facilitates the efficient and effective retention and transfer of knowledge when learning a new task. It is also a challenge for neural-symbolic integration because of the computational complexity of knowledge extraction, in general, and the need for compact representations that would enable efficient reasoning about what has been learned.
- Deep networks represent knowledge at different levels of abstraction in a modular way. This may be related to the fibring of neural networks and the representation of modal logics in neural networks, which are intrinsically modular [4, 5] and decidable, offering a sweet spot in the complexity-expressiveness landscape [15]. Modularity of deep networks seem suitable to knowledge extraction, which may help reduce the computational complexity of extraction algorithms [16], contributing to *close the gap in the pipeline* and leading to potential advances in lifelong learning, transfer learning, and applications.
- Applications: neural-symbolic computation techniques and tools have been applied effectively to action learning and knowledge description in videos [17, 18], argumentation learning in AI [19, 20], intelligent transportation systems to reduce CO_2 emissions [21], run-time verification and adaptation [22, 23], hardware/software requirements specification and verification [3, 22], normative reasoning [24], concept and ontology learning, in particular considering description logics and the semantic web [25, 26, 27], training and assessment in driving simulators, action learning and the extraction of descriptions from videos [17]. The lively demo fest organized at the seminar showed the reach of the field where promising prototypes and tools were demonstrated.

References

1 J. McCarthy. Epistemological problems for connectionism. Behav-ioral and Brain Sciences, 11:44, 1988.

2 Luís C. Lamb, Rafael V. Borges, Artur S. d'Avila Garcez: A Connectionist Cognitive Model for Temporal Synchronisation and Learning. AAAI Conference on Artificial Intelligence, AAAI-07:827–832, 2007.

3 Rafael V. Borges, Artur S. d'Avila Garcez, Luís C. Lamb: Learning and Representing Temporal Knowledge in Recurrent Networks. IEEE Transactions on Neural Networks 22(12): 2409–2421, 2011.

4 A. S. d'Avila Garcez, Luís C. Lamb, Dov M. Gabbay: Connectionist computations of intuitionistic reasoning. Theor. Comput. Sci. 358(1):34–55, 2006.

5 A. S. d'Avila Garcez, L. C. Lamb and D.M. Gabbay. Connectionist Modal Logic: Representing Modalities in Neural Networks. Theoretical Computer Science, 371(1–2):34–53, 2007.

6 S. Bader, P. Hitzler, and S. Hoelldobler, 2008. Connectionist model generation: A first-order approach. Neurocomputing, 71:2420–2432, 2008.

7 M. Kroetzsch, S. Rudolph, and P. Hitzler. Complexity of Horn de-scription logics. ACM Trans. Comput. Logic, 14(1), 2013.

8 Stephen Muggleton, Luc De Raedt, David Poole, Ivan Bratko, Peter A. Flach, Katsumi Inoue, Ashwin Srinivasan: ILP turns 20 – Biography and future challenges. Machine Learning 86(1):3–23, 2012.

9 Artur d'Avila Garcez, Tarek R. Besold, Luc de Raedt, Peter Foeldiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kuehnberger, Luis C. Lamb, Risto Miikkulainen, Daniel L. Silver. Neural-Symbolic Learning and Reasoning: Contributions and Challenges. Proceedings of the AAAI Spring Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches, Stanford, March 2015.

10 J. A. Feldman, 2006. From Molecule to Metaphor: A Neural Theory of Language, Bradford Books, Cambridge, MA: MIT Press

11 D. M. Endres and P. Foldiak. An application of formal concept analysis to semantic neural decoding. Ann. Math. Artif. Intell.57:233–248, 2009.

12 R. Khardon and D. Roth, Learning to Reason, Journal of the ACM, vol. 44, 5, pp. 697–725, 1997.

13 Valiant, L. G. Knowledge Infusion: In Pursuit of Robustness in Artificial Intelligence. FSTTCS, pp. 415–422, 2008.

14 D. Silver. The Consolidation of Task Knowledge for Lifelong Machine Learning. Proc. AAAI Spring Symposium on Lifelong Machine Learning, Stanford University, pp. 46–48, 2013.

15 M. Y. Vardi: Why is Modal Logic So Robustly Decidable? Descriptive Complexity and Finite Models, 149–184, 1996.

16 S. Tran and A. d'Avila Garcez. Knowledge Extraction from Deep Belief Networks for Images. Proc. IJCAI Workshop on Neural-Symbolic Learning and Reasoning, NeSy'13, Beijing, 2013.

17 Leo de Penning, Artur S. d'Avila Garcez, Luís C. Lamb, John-Jules Ch. Meyer: A Neural-Symbolic Cognitive Agent for Online Learning and Reasoning. IJCAI 2011:1653–1658, 2011

18 Leo de Penning, Artur S. d'Avila Garcez, John-Jules Ch. Meyer: Dreaming Machines: On multimodal fusion and information retrieval using neural-symbolic cognitive agents. ICCSW 2013:89–94, 2013

19 Artur S. d'Avila Garcez, Dov M. Gabbay, Luís C. Lamb: Value-based Argumentation Frameworks as Neural-symbolic Learning Systems. J. Log. Comput. 15(6):1041–1058, 2005.

20 A. S. d'Avila Garcez, Dov M. Gabbay, Luís C. Lamb: A neural cognitive model of argumentation with application to legal inference and decision making. J. Applied Logic 12(2):109–127, 2014.

21 Leo de Penning, Artur S. d'Avila Garcez, Luís C. Lamb, Arjan Stuiver, John-Jules Ch. Meyer: Applying Neural-Symbolic Cognitive Agents in Intelligent Transport Systems to reduce CO2 emissions. IJCNN 2014:55–62.

22 Rafael V. Borges, Artur S. d'Avila Garcez, Luís C. Lamb, Bashar Nuseibeh: Learning to adapt requirements specifications of evolving systems. ICSE 2011:856–859, 2011.

23 Alan Perotti, Artur S. d'Avila Garcez, Guido Boella: Neural Networks for Runtime Verification. IJCNN 2014:2637–2644.

24 Guido Boella, Silvano Colombo, Artur S. d'Avila Garcez, Valerio Genovese, Alan Perotti, Leendert van der Torre: Learning and reasoning about norms using neural-symbolic systems. AAMAS 2012:1023–1030.

25 C. d'Amato, Volha Bryl and Luciano Serafini. Semantic Knowledge Discovery from Heterogeneous Data Sources. In Proc. of the Knowledge Engineering and Knowledge Management – 18th International Conference (EKAW'12), vol. 7603 pp. 26–31, Springer, LNCS (2012).

26 P. Hitzler, S. Bader, and A. d'Avila Garcez, 2005. Ontology learning as a use case for neural-symbolic integration. In Proc. Workshop on Neural-Symbolic Learning and Reasoning, NeSy'05 at IJCAI-05.

27 F. A. Lisi. A declarative modeling language for concept learning in description logics. In: Riguzzi, F., Zelezny, F. (eds.) Inductive Logic Programming, 22nd International Conference, ILP 2012, Dubrovnik, Croatia, September 17–19, 2012, Revised Selected Papers. Lecture Notes in Computer Science, vol. 7842. Springer, 2013.

Participants

- Tsvi Achler
IBM Almaden Center, US
- Jim Benvenuto
HRL Labs – Malibu, US
- Tarek R. Besold
Universität Osnabrück, DE
- Srikanth Cherla
City University – London, GB
- Claudia d'Amato
University of Bari, IT
- Artur d'Avila Garcez
City University – London, GB
- James Christopher Davidson
Google Inc. –
Mountain View, US
- Leo de Penning
TNO Behaviour and Societal
Sciences – Soesterberg, NL
- Luc De Raedt
KU Leuven, BE
- Natalia Díaz-Rodríguez
Turku Centre for Computer
Science, FI
- Dominik Endres
Universität Marburg, DE
- Jacqueline Fairley
Emory University – Atlanta, US
- Jerry A. Feldman
University of California –
Berkeley, US
- Peter Földiak
University of St Andrews, GB
- Manoel Franca
City University – London, GB
- Christophe D. M. Gueret
DANS – Den Hague, NL
- Biao Han
NUDT – Hunan, CN
- Pascal Hitzler
Wright State University –
Dayton, US
- Steffen Hölldobler
TU Dresden, DE
- Thomas Icard
Stanford University, US
- Randal A. Koene
Carboncopies –
San Francisco, US
- Kai-Uwe Kühnberger
Universität Osnabrück, DE
- Luis Lamb
Federal University of Rio Grande
do Sul, BR
- Francesca Alessandra Lisi
University of Bari, IT
- Dragos Margineantu
Boeing Research & Technology –
Seattle, US
- Vivien Mast
Universität Bremen, DE
- Risto Miikkulainen
University of Texas – Austin, US
- Andrey Mokhov
Newcastle University, GB
- Bernd Neumann
Universität Hamburg, DE
- Günther Palm
Universität Ulm, DE
- Alan Perotti
University of Turin, IT
- Gadi Pinkas
Or Yehudah & Bar Ilan
University, IL
- Subramanian Ramamoorthy
University of Edinburgh, GB
- Luciano Serafini
Bruno Kessler Foundation –
Trento, IT
- Daniel L. Silver
Acadia Univ. – Wolfville, CA
- Son Tran
City University – London, GB
- Joshua Welch
University of North Carolina –
Chapel Hill, US
- Mark Wernsdorfer
Universität Bamberg, DE
- Thomas Wischgoll
Wright State University –
Dayton, US

Report from Dagstuhl Seminar 14391

Algebra in Computational Complexity

Edited by

Manindra Agrawal[1], Valentine Kabanets[2], Thomas Thierauf[3], and Christopher Umans[4]

1 **Indian Institute of Technology – Kanpur, IN,** manindra@iitk.ac.in
1 **Simon Fraser University, CA,** kabanets@cs.sfu.ca
3 **Aalen University, DE,** thomas.thierauf@htw-aalen.de
4 **CalTech – Pasadena, US,** umans@cs.caltech.edu

Abstract

At its core, much of Computational Complexity is concerned with combinatorial objects and structures. But it has often proven true that the best way to prove things about these combinatorial objects is by establishing a connection to a more well-behaved algebraic setting. Indeed, many of the deepest and most powerful results in Computational Complexity rely on algebraic proof techniques. The Razborov-Smolensky polynomial-approximation method for proving constant-depth circuit lower bounds, the PCP characterization of NP, and the Agrawal-Kayal-Saxena polynomial-time primality test are some of the most prominent examples.

The algebraic theme continues in some of the most exciting recent progress in computational complexity. There have been significant recent advances in algebraic circuit lower bounds, and the so-called “chasm at depth 4” suggests that the restricted models now being considered are not so far from ones that would lead to a general result. There have been similar successes concerning the related problems of polynomial identity testing and circuit reconstruction in the algebraic model, and these are tied to central questions regarding the power of randomness in computation. Representation theory has emerged as an important tool in three separate lines of work: the “Geometric Complexity Theory” approach to P vs. NP and circuit lower bounds, the effort to resolve the complexity of matrix multiplication, and a framework for constructing locally testable codes. Coding theory has seen several algebraic innovations in recent years, including multiplicity codes, and new lower bounds.

This seminar brought together researchers who are using a diverse array of algebraic methods in a variety of settings. It plays an important role in educating a diverse community about the latest new techniques, spurring further progress.

Seminar September 21–26, 2014 – http://www.dagstuhl.de/14391
1998 ACM Subject Classification F.1.3 Complexity Measures and Classes, F.2 Analysis of Algorithms and Problem Complexity
Keywords and phrases Computational Complexity, lower bounds, approximazation, pseudo-randomness, derandomization, circuits
Digital Object Identifier 10.4230/DagRep.4.9.85

Algebra in Computational Complexity, *Dagstuhl Reports*, Vol. 4, Issue 9, pp. 85–105
Editors: Manindra Agrawal, Valentine Kabanets, Thomas Thierauf, and Christopher Umans
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Manindra Agrawal
Valentine Kabanets
Thomas Thierauf
Christopher Umans

The seminar brought together almost 50 researchers covering a wide spectrum of complexity theory. The focus on algebraic methods showed the great importance of such techniques for theoretical computer science. We had 25 talks, most of them lasting about 40 minutes, leaving ample room for discussions. In the following we describe the major topics of discussion in more detail.

Circuit Complexity

This is an area of fundamental importance to Complexity. Circuit Complexity was one of the main topics in the seminar. Still it remains a big challenge to prove strong upper and lower bounds. However, the speakers reported amazing progress in various directions.

Or Meir talked on one of the major open problems in complexity theory: proving super-logarithmic lower bounds on the depth of circuits. That is, separating the log-depth circuit class NC^1 from polynomial time, P. Karchmer, Raz, and Wigderson suggested an approach to this problem. The *KRW-conjecture* states that the circuit depth of two functions f and g adds up when we consider the composed function $g \circ f$. They showed that the conjecture implies a separation of NC^1 from P. In his talk, Or Meir presented a natural step in this direction, which lies between what is known and the original conjecture: he showed that an analogue of the conjecture holds for the composition of a function with a universal relation. The main technical tool is to use information complexity to analyze certain communication problems.

A core theme in circuit complexity is *depth-reduction*: very roughly, these are techniques to reduce the depth of a given circuit without increasing its size too much. The classic work of Valiant, Skyum, Berkowitz and Rackoff shows that any polynomial size arithmetic circuit has an equivalent circuit of polynomial size and $\log^2 n$ depth, where n is the number of input variables. Further impedus was given by Agrawal and Vinay who pushed the depth reduction to constant depth, thereby establishing the *chasm at depth* 4. It states that exponential lower bounds for circuits of depth 4 already give such bounds for general circuits. This was further improved by Koiran and by Tavenas.

Ramprasad Saptharishi gave a slightly different proof of the depth reduction of Tavenas in his talk. Thereby he was able to apply the technique to homogeneous formulas and constant depth formulas.

Chandan Saha presented a very strong result: an exponential lower bound for homogeneous depth-4 circuits that comes close to the chasm-barrier. His techniques also yield exponential lower bounds for certain nonhomogeneous depth-3 circuits. Having the parameters so close to the bounds coming from depth reduction make these results really exciting.

Depth reduction is also an crucial ingredient in Pascal Koirans talk. He presented a new version of the τ-*conjecture* for Newton polygons of bivariate polynomials. The τ-conjecture was originally stated by Shub and Smale:

> *the number of integer roots of a univariate polynomial should be polynomially bounded in the size of the smallest straight-line program computing it.*

Pascal Koiran proposed a new version of the τ-conjecture in his talk:

> *when a bivariate polynomial is expressed as a sum of products of sparse polynomials, the number of edges of its Newton polygon is polynomially bounded in the size of such an expression.*

If this new conjecture is true, then the permanent polynomial cannot be computed by polynomial-size arithmetic circuits.

Spurred by the depth reduction results, we have seen some great work on *Polynomial Identity Testing* (PIT) recently, in particular on depth-3 and depth 4 circuits, and on arithmetic branching programs. The most ambitious goal here is to come up with a hitting set construction for a specific model. A hitting set is a set of instances such that every non-zero polynomial in the model has a non-root in the set. This solves the PIT problem in the *black box* model.

Rohit Gurjar and Arpita Korwar gave a joint talk on PIT for read-once arithmetic branching programs. They presented a new technique called *basis isolating weight assignment*. These weight assignments yield a hitting set in quasi-polynomial time.

Michael Forbes considered the question whether the hitting set constructions running in quasi-polynomial time can be improved to polynomial time. He showed that in the case of depth-3 powering circuits (sums of powers of linear polynomials) one can obtain a hitting set of size $\mathrm{poly}(s)^{\log\log s}$ for circuits of size s, which is pretty close to resolving the black-box identity testing problem for this class in polynomial time.

Swastik Kopparty showed the computational equivalence of factoring multivariate polynomials and PIT. For both problems we have efficient randomized algorithms. The question whether these algorithms can be derandomized are central in arithmetic complexity. Swastik established that they are equivalent.

Valiant introduced the arithmetic analogue of classes P and NP. Very roughly, the class VP contains all multivariate polynomials that can be computed (non-uniformly) by polynomial-size arithmetic circuits, and the class VNP contains all multivariate polynomials that have coefficients computable by VP-circuits. The question whether VP is different from VNP plays the role of the P-NP question in algebraic complexity theory. Valiant showed that the permanent is complete for VNP. But for VP, only artificially constructed functions were known to be complete. In her talk, Meena Mahajan described several natural complete polynomials for VP, based on the notion of graph homomorphism polynomials.

Eric Allender defined a class called ΛP which is in some sense dual to VP. Over finite fields, VP can be characterized by SAC^1, the class of logarithmic depth, polynomial-size semi-unbounded fan-in circuits (with bounded fan-in multiplication gates and unbounded fan-in addition gates). Eric defined the dual class ΛP in the same way, but with unbounded fan-in multiplication gates and bounded fan-in addition gates. He showed new characterizations of the complexity classes ACC^1 and TC^1 based on ΛP.

Klaus-Joern Lange defined a completeness notion on families of languages, called *densely complete*. He showed that the context-free languages are densely complete in SAC^1 via many-one AC^0-reductions.

Complexity

Ryan Williams once again demonstrated a fruitful interplay between algorithms and complexity. In his famous ACC-paper, he showed how to use fast algorithms for circuit satisfiability to prove lower bounds with respect to the class ACC. In his present talk, Ryan reversed the direction and showed how to exploit techniques from complexity to obtain faster algorithms for the all-pairs shortest paths problem (APSP). He improved the running time from $n^3/\log^2 n$ previously to $n^3/2^{\Omega(\sqrt{\log n})}$. The big question here is whether one can improve the running time to $n^{3-\epsilon}$ for some $\epsilon > 0$. A crucial role in the new algorithm plays the *polynomial method* of Razborov and Smolensky, originally conceived for proving low-depth circuit lower bounds.

Michal Koucký talked on a model of computation he calls *catalytic computation*. In this model, a machine has only limited memory available, but has additionally access to almost unlimited amount of disk space, the *catalytic memory*. This disk is however already full of data. The machine has read-write access to the disk so that it can modify the content of the disk. However, at the end of a computation, the content of the catalytic memory has to be in its original state. The question now is whether the catalytic memory is of any use. Michal showed that a logspace bounded machine with a catalytic memory can do all of nondeterministic logspace. Hence, surprisingly, the catalytic memory really helps, unless L = NL.

Amnon Ta-Shma talked on the problem of *approximating* the eigenvalues of stochastic Hermitian matrices. In an earlier paper he had shown that this is possible in probabilistic logspace in the quantum model of computation, i.e. in BQL. In this talk, Amnon was asking whether this is also possible in probabilistic logspace in the classic world, i.e. in BPL. He showed that how to achieve approximations with *constant* accuracy. To bring the problem into BPL, one would have to approximate the eigenvalues with polynomially small accuracy. This remains open for now.

Venkatesan Guruswami condidered the following promise version of the satisfiability problem: Given a k-SAT instance with the promise that there is an assignment satisfying at least t out of k literals in each clause, can one efficiently find a satisfying assignment? Because 3-SAT is NP-hard, the promise problem is NP-hard for $t \leq k/3$. On the other hand, 2-SAT is efficiently solvable. Extensions of the 2-SAT algorithm show that the promis problem is efficiently solvable for $t \geq k/2$. Venkatesan showed a sharp borderline for the promise problem: it is NP-hard for $t < k/2$. The proof uses part of the PCP-machinery.

Communication Complexity

Amir Yehudayoff talked on communication complexity in the number on the forehead model. He considered the disjointness problem: there are k players, each having a set of numbers from $[n]$. A player can see the numbers of all the other players, but not his own numbers. The task of the payers is to determine, whether there is a number common to all sets. Amir showed a lower bound for the deterministic communication complexity of order $n/4^k$. This is quite amazing since it nearly matches the known upper bound, which is of order $k^2 n/2^k$.

Arkadev Chattopadhyay talked on a communication model, where the inputs are distributed among the vertices of an undirected graph. The vertices coorespond to processors, each processor can send messages only to its neighbors in the graph. Arkadev showed lower bounds on the communication cost for computing certain functions in this model.

Rahul Santhanam considered a communication model called *compression game*. There are two players, Alice and Bob. Alice receives the whole input x and is computationally bounded, by $\mathrm{AC}^0[p]$ in this case, for some prime p. Bob has no information about x and is computationally unbounded. The communication cost of some function f is the number of bits Alice sent to Bob until they agree on the value $f(x)$. Rahul showed a lower bound on the communication complexity of the Mod_q-function, for any prime $q \neq p$.

Coding Theory

Error-correcting codes, particularly those constructed from polynomials, lie at the heart of many significant results in Computational Complexity. Usually, error correcting codes are studied with respect to the Hamming distance. Another model is that of random errors. Amir Shpilka in his talk considered the behaviour of Reed-Muller codes in the Shannon model of random errors. He showed that the rate for Reed-Muller codes with either low- or high-degree achieves (with high probability) the capacity for the Binary-Erasure-Channel

David Zuckerman talked on the relatively new concept of *non-malleable codes* which was introduced by Dziembowski, Pietrzak, and Wichs in 2010. Informally, a code is non-malleable if the message contained in a modified codeword is either the original message, or a completely unrelated value. Non-malleable codes provide an elegant algorithmic solution to the task of protecting hardware functionalities against "tampering attacks". David showed how to construct efficient non-malleable codes in the so-called C-split-state model that achieve constant rate and exponentially small error.

Game Theory

Steve Fenner considered the following two-player game on a finite partially odered set (poset) S: each player takes turns picking an element x of S and removes all $y > x$ from S. The first one to empty the poset wins. Daniel Grier showed that determining the winner of a poset game is PSPACE-complete. Steve considered the *black-white version* of the game, where each player and each element of S is assigned a color, black or white. Each player is only allowed to remove elements of their own color. He showed that also this black-white version of the poset game is PSPACE-complete. This is the first PSPACE-hardness result known for a purely numerical game. Another interesting result was that the game NimG, a generalization of both Nim and Geography, is polynomial-time solvable when restricted to undirected, bipartite graphs, whereas NimG is known to be PSPACE-complete for general graphs, both directed and undirected.

Bill Gasarch talked on a variant of classical NIM, where there is only one pile of stones and and a given set $\{a_1, a_2, \ldots, a_k\}$ of numbers. A move consists of choosing a number a_i from the set and then removing a_i stones from the pile. The first player who cannot move loses the game. This game has already been well studied. Bill considered an extension of the game where each player starts out with a number of dollars. Now each player has to spend a dollars to remove a stones. He presented some surprising results on the winning conditions for the extended game.

Cryptography

Farid Ablayev generalized classical universal hashing to the quantum setting. He defined the concept of a quantum hash generator and offer a design, which allows one to build a large number of different quantum hash functions. One of the important points here is to use unly few quantum bits. Farid proved that his construction is optimal with respect to the number of qubits needed.

Matthias Krause talked on approaches for designing authentication protocols for ultralight weight devices as for example RFID chips. He proposed a new approach based on key stream generators as the main building block.

Conclusion

As is evident from the list above, the talks ranged over a broad assortment of subjects with the underlying theme of using algebraic and combinatorial techniques. It was a very fruitful meeting and has hopefully initiated new directions in research. Several participants specifically mentioned that they appreciated the particular focus on a common class of *techniques* (rather than end results) as a unifying theme of the workshop. We look forward to our next meeting!

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 Quantum hashing via classical ϵ-universal hashing constructions

Farid Ablayev (Kazan State University, RU)

Joint work of Farid Ablayev, Marat Ablayev

Quantum computing is inherently a very mathematical subject, and discussions of how quantum computers can be more efficient than classical computers in breaking encryption algorithms have started since Peter Shor invented his famous quantum algorithm. The reaction of a cryptography community is a "Post-quantum cryptography", which refers to the research of problems (usually public-key cryptosystems) that are not efficiently breakable using quantum computers. Currently post-quantum cryptography includes different approaches, in particular, hash-based signature schemes such as Lamport signature and Merkle signature scheme.

Hashing itself is an important basic concept of computer science. The concept known as "universal hashing" was invented by Carter and Wegman in 1979.

In our research we define a quantum hashing as a quantum generalization of classical hashing. We define the concept of a quantum hash generator and offer a design, which allows one to build a large number of different quantum hash functions. The construction is based on composition of a classical ϵ-universal hash family and a given family of functions – quantum hash generators.

The relationship between epsilon-universal hash families and error-correcting codes give possibilities to build a large amount of different quantum hash functions. In particular, we present quantum hash function based on Reed-Solomon code, and we proved, that this construction is optimal in the sense of number of qubits needed.

Using the relationship between epsilon-universal hash families and Freivalds' fingerprinting schemas we present explicit quantum hash function and prove that this construction is optimal with respect to the number of qubits needed for the construction.

3.2 Dual VP classes

Eric Allender (Rutgers University, US)

Joint work of Eric Allender, Anna Gal, Ian Mertz
Main reference E. Allender, A. Gal, I. Mertz, "Dual VP Classes," ECCC, TR14-122, 2014.
URL http://eccc.hpi-web.de/report/2014/122/

We consider arithmetic complexity classes that are in some sense dual to the classes VP that were introduced by Valiant. This provides new characterizations of the complexity classes ACC^1 and TC^1, and also provides a compelling example of a class of high-degree polynomials that can be simulated via arithmetic circuits of much lower degree.

3.3 Asymptotic spectra of tensors

Markus Bläser (Universität des Saarlandes, DE)

Joint work of Manuel Arora, Markus Bläser

Asymptotic spectra were studied by Strassen to understand the asymptotic complexity of tensors, in particular of matrix multiplication. The (equivalence classes of) tensors are embedded into an ordered ring and then results by Stone, Kadison, and Dubois are applied to represent tensors by nonnegative continuous functions on some Hausdorff space.

In the first part of the talk, we give an introduction to asymptotic spectra and the work by Strassen. In the second part of the talk, we introduce a new order on the equivalence classes of tensors and study the resulting new spectra.

3.4 Topology matters in communication

Arkadev Chattopadhyay (TIFR, IN)

Joint work of Arkadev Chattopadhyay, Jaikumar Radhakrishnan, Atri Rudra

We consider the communication cost of computing functions when inputs are distributed among the vertices of an undirected graph. The communication is assumed to be point-to-point: a processor sends messages only to its neighbors. The processors in the graph act according to a pre-determined protocol, which can be randomized and may err with some small probability. The communication cost of the protocol is the total number of bits exchanged in the worst case. Extending recent work that assumed that the graph was the complete graph (with unit edge lengths), we develop a methodology for showing lower bounds that are sensitive to the graph topology. In particular, for a broad class of graphs, we obtain a lower bound of the form k^2n, for computing a function of k inputs, each of which is n-bits long and located at a different vertex. Previous works obtained lower bounds of the form kn.

This methodology yields a variety of other results including the following:

- A tight lower bound (ignoring poly-log factors) for Element Distinctness, settling a question of Phillips, Verbin and Zhang (SODA'12);
- a distributed XOR lemma;
- a lower bound for composed functions, settling a question of Phillips et al.;
- new topology-dependent bounds for several natural graph problems considered by Woodruff and Zhang (DISC'13).

To obtain these results we use tools from the theory of metric embeddings and represent the topological constraints imposed by the graph as a collection of cuts, each cut providing a setting where our understanding of two-party communication complexity can be effectively deployed.

3.5 Some new results on combinatorial game complexity

Stephen A. Fenner (University of South Carolina, US)

Joint work of Daniel Grier, Stephen A. Fenner Daniel, Jochen Messner, Luke Schaeffer, Thomas Thierauf

We give new hardness and easiness results for determining the winner in certain two-player games with perfect information. On the hardness side, we show that Black-White-Poset-Games (BWPG) and a generalized version of the game Col are both PSPACE-complete (via reductions from variants of TQBF). The BWPG result is the first PSPACE-hardness result known for a purely numerical game. On the easiness side, we show that NimG (a generalization of both Nim and Geography) is polynomial-time computable when restricted to undirected, bipartite graphs. (NimG is known to be PSPACE-complete for general graphs, both directed and undirected). We also show that Toads and Frogs is polynomial-time computable when each row is restricted to one toad and one frog.

3.6 Hitting Sets for Depth-3 Powering Circuits

Michael Forbes (University of California – Berkeley, US)

Joint work of Michael Forbes, Ramprasad Saptharishi, Amir Shpilka

A recent line of research has constructed hitting sets for various read-once and set-multilinear models of computation, as such hitting sets yield black-box polynomial identity testing algorithms. Despite the fact that these models all have a "white-box" identity testing algorithm that runs in polynomial-time (due to Raz and Shpilka), the black-box algorithms all run in quasipolynomial time. Improving these algorithms seems challenging, especially as these algorithms can be viewed as algebraic analogues of pseudorandom generators for RL (which have been stuck at $RL \subset L^2$ for 25 years).

In this work, we identify a particularly simple subclass of the above models, known as depth-3 powering circuits (sums of powers of linear polynomials). In fact, this is the simplest complete algebraic circuit class for which we do not have explicit polynomial-size hitting sets. We show how to combine two different hitting set constructions, each of size $\mathrm{poly}(s)^{\log s}$ for size s circuits, to obtain a hitting set of size $\mathrm{poly}(s)^{\log \log s}$, which is tantalizingly close to resolving the black-box identity testing problem for this class.

3.7 NIM with Cash

William Gasarch (University of Maryland – College Park, US)

Joint work of William Gasarch, John Purtilo, Doug Ulrich
Main reference W. Gasarch, J. Purtilo, "NIM with Cash," University of Maryland Computer Science Department, CS-TR-5015, 2012.
URL http://hdl.handle.net/1903/12908

NIM$(a_1, \ldots, a_k; n)$ is a 2-player game where initially there are n stones on the board and the players alternate removing either a_1 or $\ldots$ a_k stones. The first player who cannot move loses. This game has been well studied. For example, it is known that for NIM$(1, 2, 3; n)$

Player II wins if and only if n is divisible by 4. This game is interesting because even small sets $\{a_1, \ldots, a_k\}$ lead to interesting win conditions.

We investigate an extension of the game where Player I starts out with d_1 dollars and Player II starts out with d_2 dollars, and a player has to spend a dollars to remove a stones. For several choices of $a_1, \ldots, a_k$ we determine for all (n, d_1, d_2) which player wins. The win condition depend on *both* what n is congruent to mod some M *and* on how d_1 and d_2 relate. This game is interesting because even small sets $\{a_1, \ldots, a_k\}$ lead to interesting and complicated win conditions.

Some of our results are surprising. For example, there are cases where both players are poor, yet the one with less money wins.

3.8 Hitting Set for Read-Once Arithmetic Branching Programs

Rohit Gurjar and Arpita Korwar (IIT Kanpur, IN)

Joint work of Manindra Agrawal, Rohit Gurjar, Arpita Korwar, Nitin Saxena
Main reference M. Agrawal, R. Gurjar, A. Korwar, N. Saxena, "Hitting-sets for ROABP and Sum of Set-Multilinear circuits," ECCC, TR14-085, 2014.
URL http://eccc.hpi-web.de/report/2014/085/

In the march towards a deterministic solution for the polynomial identity testing problem, recently there has been a considerable amount of work on depth-3 set-multilinear circuits and read once arithmetic branching programs (ROABP). Continuing in this direction, we have given a $(n\delta)^{O(\log n)}$-time blackbox PIT algorithm for unknown-order, n-variate, individual degree δ ROABP, improving the previously known $n^{O(\delta \log^2 n)}$-time algorithm.

In this talk, we will look at a new idea "Basis Isolating Weight Assignment" for designing a hitting set for depth-3 circuits. This idea has been applied to read-once arithmetic branching programs (RO-ABPs) to get a $n^O(\log n)$ time hitting set.

3.9 (2+eps)-SAT is NP-hard

Venkatesan Guruswami (Carnegie Mellon University , US)

Joint work of Per Austrin, Venkatesan Guruswami, Johan Håstad
Main reference P. Austrin, V. Guruswami, J. Håstad, "$(2+\epsilon)$-SAT is NP-hard," ECCC, TR13-159, 2013/2014; peer-reviewed version to appear in Proc. of the 55th IEEE Symp. on Foundations of Computer Science (FOCS'14).
URL http://eccc.hpi-web.de/report/2013/159/

Given a k-SAT instance with the promise that there is an assignment satisfying at least t out of k literals in each clause, can one efficiently find a satisfying assignment (setting at least one literal to true in every clause)? The NP-hardness of 3-SAT implies that this problem is NP-hard when $t \leq k/3$, and extensions of some 2-SAT algorithms give efficient solutions when $t \geq k/2$.

We prove that for $t < k/2$, the problem is NP-hard. Thus, satisfiability becomes hard when the promised density of true literals falls below 1/2. One might thus say that the transition from easy to hard in 2-SAT vs. 3-SAT takes place just after two and not just before three.

The talk will sketch most of the proof, which is based on the fact that the only functions passing a natural dictatorship test are "juntas" depending on few variables. We will briefly mention the general "universal-algebraic" principle (based on the lack of certain *polymorphisms*) that underlies hardness of constraint satisfaction.

A strengthening of the k-SAT result shows that given a $(2t+1)$-uniform hypergraph that can be 2-colored such that each edge has near-perfect balance (at most $t+1$ vertices of each color), it is NP-hard to even find a 2-coloring that avoids a monochromatic edge. This shows extreme hardness of discrepancy minimization for systems of bounded-size sets.

(Subsequent work with Euiwoong Lee, available as ECCC TR14-043 and to appear at SODA 2015, in fact rules out coloring with any constant number of colors for the case of $2k$-uniform hypergraphs with discrepancy 2, and shows further extensions to hypergraphs admitting a near-balanced rainbow coloring with more than two colors.)

3.10 A τ-conjecture for Newton polygons

Pascal Koiran (ENS – Lyon, FR)

Joint work of Pascal Koiran, Natacha Portier, Sébastien Tavenas, Stéphan Thomassé
Main reference P. Koiran, N. Portier, S. Tavenas, S. Thomassé, "A tau-conjecture for Newton polygons," arXiv:1308.2286v2 [cs.CC], 2014.
URL http://arxiv.org/abs/1308.2286v2

One can associate to any bivariate polynomial $P(X, Y)$ its Newton polygon. This is the convex hull of the points (i, j) such that the monomial X^iY^j appears in P with a nonzero coefficient. We conjecture that when P is expressed as a sum of products of sparse polynomials, the number of edges of its Newton polygon is polynomially bounded in the size of such an expression. We show that this "τ-conjecture for Newton polygons," even in a weak form, implies that the permanent polynomial is not computable by polynomial size arithmetic circuits. We make the same observation for a weak version of an earlier "real τ-conjecture." Finally, we make some progress toward the τ-conjecture for Newton polygons using recent results from combinatorial geometry.

3.11 Equivalence of polynomial identity testing and multivariate polynomial factorization

Swastik Kopparty (Rutgers University, US)

Joint work of Swastik Kopparty, Shubhangi Saraf, Amir Shpilka
Main reference S. Kopparty, S. Saraf, A. Shpilka, "Equivalence of Polynomial Identity Testing and Deterministic Multivariate Polynomial Factorization," ECCC, TR14-001, 2014.
URL http://eccc.hpi-web.de/report/2014/001/

In this work, we show that the problem of deterministically factoring multivariate polynomials reduces to the problem of deterministic polynomial identity testing. Specifically, we show that given an arithmetic circuit (either explicitly or via black-box access) that computes a polynomial $f(X_1, \ldots, X_n)$, the task of computing arithmetic circuits for the factors of f can be solved deterministically, given a deterministic algorithm for the polynomial identity testing problem (we require either a white-box or a black-box algorithm, depending on the representation of f).

Together with the easy observation that deterministic factoring implies a deterministic algorithm for polynomial identity testing, this establishes an equivalence between these two central derandomization problems of arithmetic complexity. Previously, such an equivalence was known only for multilinear circuits (Shpilka and Volkovich, ICALP 2010).

3.12 Catalytic computation

Michal Koucký (Charles University, CZ)

Joint work of Harry Buhrman, Richard Cleve, Michal Koucký, Bruno Loff, Florian Speelman
Main reference H. Buhrman, R. Cleve, M. Koucky, B. Loff, F. Speelman, "Computing with a full memory: Catalytic space," ECCC, TR14-053, 2014.
URL http://www.eccc.hpi-web.de/report/2014/053/

The known hierarchy theorems hold in a vacuum. However, our computation happens in a wider context. Although we may have only limited memory to carry out our computation we have access to almost unlimited amount of disk space provided at the end of the computation the disk contains exactly the same content as at the beginning. This naturally leads to a question: what can be computed in space s when we have access to read-write "catalytic" memory that we can use provided at the end of the computation the content of the catalytic memory is at its original, possibly incompressible, state. Is there any advantage in having this extra catalytic memory?

We provide affirmative answer to this question (assuming NL differs from L). We show that in space s with catalytic memory we can compute deterministically functions computable in non-deterministic space s. We can extend the results even further. The main techniques come from a special form of reversible computation that we call transparent computation.

3.13 Sharp Security Bounds for Authentication with Key Stream Generators

Matthias Krause (Mannheim University, DE)

In the last years, various approaches for designing authentication protocols for ultralight weight devices (e.g., RFIDs) have been intensively studied (HB-type protocols, Linear protocols, block cipher based solutions etc.) We propose an new approach which uses a key stream generators (KSG) as the main building block. The usage of KSGs appears advantageous in this context, as several well analyzed ultralight weight practical designs are available.

We propose a new mode of operation for KSGs which leads to an encryption function $E = E(x)$ of type $E(x) = F(P(x + k_1) + k_2)$, where F denotes a pseudo-random function, P a pseudo-random permutation and k_1, k_2 secret keys of length n.

We show a sharp information theoretic bound of $\frac{2}{3}n$ for the effective key length of this construction w.r.t. to an attacker of unbounded computational power which has access to E-, F- and P, P^{-1}-oracles.

3.14 Dense Completeness

Klaus-Joern Lange (Universität Tübingen, DE)

A family of formal languages F is said to be *densely complete* in a complexity class $\mathcal{C}$, iff F is contained in $\mathcal{C}$ and for each $L \in \mathcal{C}$ there exists some $L' \in F$ such that both L is reducible to L' and L' is reducible to L, i.e., L and L' have the same complexity modulo the chosen notion of reducibility.

Using many-one reductions computable in AC^0, it can be shown that the context-free languages are densely complete in SAC^1, the one-counter languages in $\mathrm{Nspace}(\log n)$, and the indexed languages in NP. On the other, hand the regular languages are not densely complete in NC^1. This result is now extended to the nonregular family of visibly one-counter languages.

3.15 Homomorphism polynomials complete for VP

Meena Mahajan (The Institute of Mathematical Sciences – Chennai, IN)

Joint work of Arnaud Durand, Meena Mahajan, Guillaume Malod, Nicolas de Rugy-Altherre, Nitin Saurabh

The VP versus VNP question, introduced by Valiant, is probably the most important open question in algebraic complexity theory. Thanks to completeness results, a variant of this question, VBP versus VNP, can be succintly restated as asking whether the permanent of a generic matrix can be written as a determinant of a matrix of polynomially bounded size. Strikingly, this restatement does not mention any notion of computational model. To get a similar restatement for the original and more fundamental question, and also to better understand the class itself, we need a complete polynomial for VP. Ad hoc constructions yielding complete polynomials were known, but not natural examples in the vein of the determinant. This talk describes several variants of natural complete polynomials for VP, based on the notion of graph homomorphism polynomials.

3.16 Toward Better Formula Lower Bounds: An Information Complexity Approach to the KRW Composition Conjecture

Or Meir (Institute of Advanced Study – Princeton, US)

Joint work of Dmitry Gavinsky, Or Meir, Omri Weinstein, Avi Wigderson
Main reference D. Gavinsky, O. Meir, O. Weinstein, A. Wigderson, “Toward Better Formula Lower Bounds: An Information Complexity Approach to the KRW Composition Conjecture,” ECCC, TR13-190, 2013/2014.
URL http://www.eccc.hpi-web.de/report/2013/190/

One of the major open problems in complexity theory is proving super-logarithmic lower bounds on the depth of circuits (i.e., $\mathbf{P} \not\subseteq \mathbf{NC}^1$). This problem is interesting for two reasons:

first, it is tightly related to understanding the power of parallel computation and of small-space computation; second, it is one of the first milestones toward proving super-polynomial circuit lower bounds.

Karchmer, Raz, and Wigderson suggested to approach this problem by proving the following conjecture: given two boolean functions f and g, the depth complexity of the composed function $g \circ f$ is roughly the sum of the depth complexities of f and g. They showed that the validity of this conjecture would imply that $\mathbf{P} \not\subseteq \mathbf{NC}^1$.

As a starting point for studying the composition of functions, they introduced a relation called ?the universal relation?, and suggested to study the composition of universal relations. This suggestion proved fruitful, and an analogue of the KRW conjecture for the universal relation was proved by Edmonds et. al. An alternative proof was given later by Håstad and Wigderson. However, studying the composition of functions seems more difficult, and the KRW conjecture is still wide open.

In this work, we make a natural step in this direction, which lies between what is known and the original conjecture: we show that an analogue of the conjecture holds for the composition of a function with a universal relation. We also suggest a candidate for the next step and provide initial results toward it.

Our main technical contribution is developing an approach based on the notion of information complexity for analyzing KW relations – communication problems that are closely related to questions on circuit depth and formula complexity. Recently, information complexity has proved to be a powerful tool, and underlined some major progress on several long-standing open problems in communication complexity. In this work, we develop general tools for analyzing the information complexity of KW relations, which may be of independent interest.

3.17 A Geometric Resolution-based Framework for Joins

Atri Rudra (SUNY – Buffalo, US)

Joint work of Mahmoud Abo Khamis, Hung Ngo, Dung Nguyen, Chris Re, Atri Rudra

We present a simple geometric framework for the relational join. Using this framework, we design an algorithm that achieves the fractional hypertree-width bound, which generalizes classical and recent worst-case algorithmic results on computing joins. In addition, we use our framework and the same algorithm to show a series of what are colloquially known as beyond worst-case results. The framework allows us to prove results for data stored in Btrees, multidimensional data structures, and even multiple indices per table. A key idea in our framework is formalizing the inference one does with an index as a type of geometric resolution; transforming the algorithmic problem of computing joins to a geometric problem. Our notion of geometric resolution can be viewed as a geometric analog of logical resolution.

In this talk, I will focus on our geometric interpretation of joins and give a flavor of our beyond worst-case results. In particular, I will present the main (very simple!) algorithmic ideas behind our upper bounds and clarify the actual model of resolution that we use. I will end with some open questions on lower bounds and some algebraic versions of the join problem that we do not know much about.

3.18 Lower bounds for (homogeneous) depth-4 and (nonhomogeneous) depth-3 arithmetic circuits

Chandan Saha (Indian Institute of Science – Bangalore, IN)

Joint work of Neeraj Kayal, Chandan Saha, Srikanth Srinivasan, Nutan Limaye

An approach to proving a super-polynomial lower bound for arithmetic circuits reduces the problem to proving "strong enough" lower bounds for small depth circuits, in particular (nonhomogeneous) depth-3 circuits and (homogeneous) depth-4 circuits. Depth of a circuit is the number of layers of gates in it.

In the talk, we plan to discuss an exponential lower bound for (homogeneous) depth-4 circuits that comes close to being 'strong enough'. More precisely, we give an explicit family of polynomials of degree d on N variables (with $N = d^3$ in our case) with $0, 1$-coefficients such that for any representation of a polynomial f in this family of the form

$$f = \sum_i \prod_j Q_{ij},$$

where the Q_{ij}'s are homogeneous polynomials (recall that a polynomial is said to be homogeneous if all its monomials have the same degree), it must hold that

$$\sum_{i,j} (\text{Number of monomials of } Q_{ij}) \quad \geq \quad 2^{\Omega(\sqrt{d}\cdot\log N)}.$$

The above mentioned family, which we refer to as the Nisan-Wigderson design-based family of polynomials, is in the complexity class VNP. Our work builds on several recent lower bound results and the techniques also yield exponential lower bounds for certain (nonhomogeneous) depth-3 circuits, in particular depth-3 circuits with low bottom fanin which also answers a question posed by Shpilka and Wigderson (CCC'99).

3.19 Lower Bounds on $\mathrm{AC}^0[p]$-Compression Games

Rahul Santhanam (University of Edinburgh, GB)

Joint work of Igor Carboni Oliveira, Rahul Santhanam

Given a class of circuits $\mathcal{C}$, a $\mathcal{C}$-compression game to compute a Boolean function f is a 2-player game played as follows. Alice is a computationally bounded player who receives the input x, and whose next-message function is computable in $\mathcal{C}$. Bob is a computationally unbounded player who has no information about x before communication happens. Alice and Bob communicate until they agree on the value of $f(x)$. The cost of a compression protocol is the number of bits communicated from Alice to Bob. Compression games hybridize computational complexity and communication complexity. They generalize the notion of instance compression due to Harnik & Naor and Bodlaender, Downey, Fellows & Hermelin, and have applications in cryptography, parameterized complexity and circuit complexity.

We prove new lower bounds for $\mathcal{C}$-compression games where $\mathcal{C} = \mathrm{AC}^0[p]$ for some prime p. We show that the Mod_q function requires deterministic compression cost $\Omega(n/\mathrm{polylog}(n))$, and randomised compression cost $\Omega(\sqrt{n}/\mathrm{polylog}(n))$, whenever q is a prime different from p.

We also define and study multi-player compression games, where Alice communicates in parallel with several unbounded players $\text{Bob}_1, \text{Bob}_2, \ldots, \text{Bob}_k$ (which cannot communicate with each other), and the cost of the protocol is the maximum amount of communication from Alice to any fixed Bob_i. We show compression cost lower bound $n^{\Omega(1)}$ for constant-round multi-player $\text{AC}^0[p]$-compression games computing the Mod_q function when $q \neq p$, even when $k = \text{poly}(n)$. As an application, we strengthen the known $\text{AC}^0[p]$ lower bounds of Razborov and Smolensky to the setting of oracle circuits with arbitrary oracle gates, with some mild restrictions on the number of layers and fan-in of the oracle gates.

Finally we obtain a stronger version of the round separation result of Chattopadhyay & Santhanam for AC^0-compression games.

3.20 Depth Reduction for Arithmetic Circuits

Ramprasad Saptharishi (Microsoft Research India – Bangalore, IN)

Joint work of Saptharishi, Ramprasad; Vinay, V.

Almost all attempts to prove lower bounds for subclasses arithmetic circuits proceed by addressing a "depth four analogue" of the subclass. This talk shall give a slightly different proof of the depth reduction of Tavenas, and enable us to study this for homogeneous formulas and constant depth formulas.

3.21 Reed-Muller codes with respect to random errors and erasures

Amir Shpilka (Technion – Haifa, IL)

Joint work of Emmanuel Abbe, Amir Shpilka, Avi Wigderson

In TCS we usually study error correcting codes with respect to the Hamming metric, i.e. we study their behaviour with respect to worst case errors. However, in coding theory a more common model is that of random errors, where Shannon's results show a much better tradeoff between rate and decoding radius.

We consider the behaviour of Reed-Muller codes in the Shannon model of random errors. In particular, we show that RM codes with either low- or high-degree (degree $n^{1/2}$ or $n - n^{1/2}$, respectively), with high probability, can decode from an $1 - R$ fraction of random erasures (where R is the rate). In other words, for this range of parameters RM codes achieve capacity for the Binary-Erasure-Channel. This result matches experimental observations that RM codes can achieve capacity for the BEC, similarly to Polar codes. We also show that RM-codes can handle many more random errors than the minimum distance, i.e. roughly $n^{r/2}$ errors for codes of degree $n - r$ (where the minimum distance is only 2^r).

We show that the questions regarding the behaviour of Reed-Muller codes wrt random errors are tightly connected to the following question. Given a random set of vectors in $\{0,1\}^n$, what is the probability the their r^{th} tensor products are linearly independent? We obtain our results by giving answer to this question for certain range of parameters.

3.22 On the problem of approximating the eigenvalues of undirected graphs in probabilistic logspace

Amnon Ta-Shma (Tel Aviv University, IL)

We focus on the problem of *approximating* the eigenvalues of stochastic Hermitian operators in small space, which is a natural and important problem. The ultimate goal is solving the problem in full in BPL, i.e., with polynomially-small accuracy. In this paper, however, we only achieve approximations with *constant* accuracy. Our technique is new. We also show that going beyond constant accuracy requires a new idea.

3.23 Faster All-Pairs Shortest Paths Via Circuit Complexity

Ryan Williams (Stanford University, US)

I presented an algorithm for solving the all-pairs shortest paths problem on n-node graphs with edge weights in $[0, n^k]$ (for arbitrary k) running in $n^3/2^{(\log n)^\delta}$ time for an unspecified $\delta > 0$. In the full paper, I give an algorithm for solving the all-pairs shortest paths problem on n-node real-weighted graphs in the "real RAM" model, running in $n^3/2^{\Omega(\sqrt{\log n})}$ time.

Both algorithms apply the *polynomial method* of Razborov and Smolensky, originally conceived for proving low-depth circuit lower bounds. We show how low-depth circuits can compute a so-called "min-plus inner product" of two vectors, then show how to evaluate such low-depth circuits efficiently on many pairs of vectors by randomly reducing the circuit to a low-degree polynomial over $\mathbb{F}_2$ and using fast rectangular matrix multiplication.

3.24 Lower bounds on the multiparty communication complexity of disjointness

Amir Yehudayoff (Technion – Haifa, IL)

We give a proof of order $n/4^k$ lower bound for the deterministic communication complexity of set disjointness with k players in the number on the forehead model. This is the first lower bound that is linear in n, and it nearly matches the known upper bound. We discuss Sherstov's proof of an order $n^{1/2}/(k2^k)$ lower bound on the randomized complexity.

3.25 Non-Malleable Codes Against Constant Split-State Tampering

David Zuckerman (University of Texas at Austin, US)

Joint work of Eshan Chattopadhyay, David Zuckerman
Main reference E. Chattopadhyay, D. Zuckerman, "Non-Malleable Codes Against Constant Split-State Tampering," ECCC, TR14-102, 2014.
URL http://eccc.hpi-web.de/report/2014/102/

Non-malleable codes were introduced by Dziembowski, Pietrzak and Wichs as an elegant generalization of the classical notion of error detection, where the corruption of a codeword is viewed as a tampering function acting on it. Informally, a non-malleable code with respect to a family of tampering functions $\mathcal{F}$ consists of a randomized encoding function Enc and a deterministic decoding function Dec such that for any m, $\mathrm{Dec}(\mathrm{Enc}(m)) = m$. Further, for any tampering function $f \in \mathcal{F}$ and any message m, $\mathrm{Dec}(f(\mathrm{Enc}(m)))$ is either m or is ϵ-close to a distribution D_f independent of m, where ϵ is called the error.

Of particular importance are non-malleable codes in the C-split-state model. In this model, the codeword is partitioned into C equal sized blocks and the tampering function family consists of functions $(f_1, \ldots, f_C)$ such that f_i acts on the i^{th} block. For $C = 1$ there cannot exist non-malleable codes. For $C = 2$, the best known explicit construction is by Aggarwal, Dodis and Lovett who achieve rate $= \Omega(n^{-6/7})$ and error $= 2^{-\Omega(n^{-1/7})}$, where n is the block length of the code.

In our main result, we construct efficient non-malleable codes in the C-split-state model for $C = 10$ that achieve constant rate and error $= 2^{-\Omega(n)}$. These are the first explicit codes of constant rate in the C-split-state model for any $C = o(n)$, that do not rely on any unproven assumptions. We also improve the error in the explicit non-malleable codes constructed in the bit tampering model by Cheraghchi and Guruswami.

Our constructions use an elegant connection found between seedless non-malleable extractors and non-malleable codes by Cheraghchi and Guruswami. We explicitly construct such seedless non-malleable extractors for 10 independent sources and deduce our results on non-malleable codes based on this connection. Our constructions of extractors use encodings and a new variant of the sum-product theorem.

Participants

- Farid Ablayev
Kazan State University, RU
- Manindra Agrawal
IIT – Kanpur, IN
- Eric Allender
Rutgers Univ. – Piscataway, US
- Vikraman Arvind
The Institute of Mathematical Sciences, IN
- Markus Bläser
Universität des Saarlandes, DE
- Andrej Bogdanov
Chinese Univ. of Hong Kong, HK
- Harry Buhrman
CWI – Amsterdam, NL
- Sourav Chakraborty
Chennai Mathematical Inst., IN
- Arkadev Chattopadhyay
TIFR Mumbai, IN
- Stephen A. Fenner
University of South Carolina – Columbia, US
- Michael Forbes
University of California – Berkeley, US
- Lance Fortnow
Georgia Inst. of Technology, US
- Anna Gál
University of Texas – Austin, US
- William Gasarch
University of Maryland, US
- Frederic Green
Clark University – Worcester, US
- Rohit Gurjar
IIT – Kanpur, IN
- Venkatesan Guruswami
Carnegie Mellon University, US
- Valentine Kabanets
Simon Fraser University – Burnaby, CA
- Marek Karpinski
Universität Bonn, DE
- Neeraj Kayal
Microsoft Research India – Bangalore, IN
- Pascal Koiran
ENS – Lyon, FR
- Swastik Kopparty
Rutgers Univ. – Piscataway, US
- Arpita Korwar
IIT – Kanpur, IN
- Michal Koucký
Charles University – Prague, CZ
- Matthias Krause
Universität Mannheim, DE
- Klaus-Jörn Lange
Universität Tübingen, DE
- Sophie Laplante
University Paris-Diderot, FR
- Meena Mahajan
The Institute of Mathematical Sciences, IN
- Or Meir
Institute of Advanced Study – Princeton, US
- Peter Bro Miltersen
Aarhus University, DK
- Natacha Portier
ENS – Lyon, FR
- Atri Rudra
SUNY – Buffalo, US
- Chandan Saha
Indian Institute of Science – Bangalore, IN
- Rahul Santhanam
University of Edinburgh, GB
- Ramprasad Saptharishi
Microsoft Research India – Bangalore, IN
- Uwe Schöning
Universität Ulm, DE
- Ronen Shaltiel
University of Haifa, IL
- Amir Shpilka
Technion – Haifa, IL
- Florian Speelman
CWI – Amsterdam, NL
- Amnon Ta-Shma
Tel Aviv University, IL
- Thomas Thierauf
Hochschule Aalen, DE
- Jacobo Torán
Universität Ulm, DE
- Christopher Umans
CalTech, US
- Nikolay K. Vereshchagin
Moscow State University, RU
- Ryan Williams
Stanford University, US
- Amir Yehudayoff
Technion – Haifa, IL
- David Zuckerman
University of Texas – Austin, US

Report from Dagstuhl Perspectives Workshop 14401

Privacy and Security in an Age of Surveillance

Edited by

Bart Preneel[1], Phillip Rogaway[2], Mark D. Ryan[3], and Peter Y. A. Ryan[4]

1 **KU Leuven and iMinds, BE,** `Bart.Preneel@esat.kuleuven.be`
2 **University of California, Davis, US,** `rogaway@cs.ucdavis.edu`
3 **University of Birmingham, GB,** `m.d.ryan@cs.bham.ac.uk`
4 **University of Luxembourg, LU,** `peter.ryan@uni.lu`

Abstract

The Snowden revelations have demonstrated that the US and other nations are amassing data about people's lives at an unprecedented scale. Furthermore, these revelations have shown that intelligence agencies are not only pursuing passive surveillance over the world's communication systems, but are also seeking to facilitate such surveillance by undermining the security of the internet and communications technologies. Thus the activities of these agencies threatens not only the rights of individual citizens but also the fabric of democratic society.

Intelligence services do have a useful role to play in protecting society and for this need the capabilities and authority to perform targeted surveillance. But the scope of such surveillance must be strictly limited by an understanding of its costs as well as benefits, and it should not impinge on the privacy rights of citizens any more than necessary.

Here we report on a recent Dagstuhl Perspectives Workshop addressing these issues – a four-day gathering of experts from multiple disciplines connected with privacy and security. The meeting explored the scope of mass-surveillance and the deliberate undermining of the security of the internet, defined basic principles that should underlie needed reforms, and discussed the potential for technical, legal and regulatory means to help restore the security of the internet and stem infringement of human-rights by ubiquitous electronic surveillance.

Perspectives Workshop September 28 to October 2, 2014 – http://www.dagstuhl.de/14401
1998 ACM Subject Classification E.3 Data Encryption, K.4.1 Public Policy Issues, K.4.2 Social Issues, K.5.2 Governmental Issues, K.6.4 Security and Protection
Keywords and phrases Big data, encryption, mass surveillance, privacy
Digital Object Identifier 10.4230/DagRep.4.9.106

1 Executive Summary

Bart Preneel
Phillip Rogaway
Mark D. Ryan
Peter Y. A. Ryan

Revelations over the last few years have made clear that the world's intelligence agencies surveil essentially everyone, recording and analyzing who you call, what you do on the web, what you store in the cloud, where you travel, and more. Furthermore, we have learnt that intelligence agencies intentionally subvert security protocols. They tap undersea cables.

They install malware on an enormous number targets worldwide. They use active attacks to undermine our network infrastructure. And they use sophisticated analysis tools to profile individuals and groups.

While we still understand relatively little about who is doing what, the documents leaked by Snowden have led to the conclusion that the Five Eyes[1] organizations are going far beyond anything necessary or proportionate for carrying legitimate intelligence activities. ot an equivalent access to documents Governmental assurances of oversight have come to ring hollow, as any oversight to date seems to have been ineffectual, and is perhaps a complete sham.

Can democracy or nonconformity survive if the words and deeds of citizens are to be obsessively observed by governments and their machines? The rise of electronic surveillance thus raises questions of immense significance to modern society. There is an inherent tension. Machine-monitored surveillance of essentially everything people do is now possible. And there are potential economic, political, and safety benefits that power may reap if it can implement effective population-wide surveillance. But there is also a human, social, economic, and political harm that can spring from the very same activity.

The goal of our workshop was to gather together a mix of people with knowledge and expertise in both the legal and technological aspects of privacy and surveillance, to try to understand the landscape that we now live in, and to debate approaches to moving forward. We invited people from a wide range of domains, including members of the intelligence community. All invitees in the intelligence community declined the invitations – in most cases choosing not even to reply. Also, we found that we had more success in getting positive replies from members of the technical community than members of the legal or regulatory communities. Consequently, the makeup of the workshop was not as diverse and balanced as we had hoped. Nonetheless, we felt that we achieved a healthy mix, and there was plenty of lively debate. The issues addressed by this workshop were unusually contentious, and discussions at times were highly animated, even heated.

It is often argued that privacy is not an absolute right. This is true, but this is also true of other rights. The right to freedom must be tempered by the fact that people who are convicted of crimes may forfeit this right for a period. Equally, someone for whom there are sound grounds for suspicion might forfeit some privacy rights. But in any event, any such breaches must be targeted and proportionate and justified by well-founded grounds for suspicion.

An important observation that came up repeatedly in discussions is that privacy is not just an individual right but essential to the health of a democratic society as a whole.

How can society as whole be provided strong assurance that intelligence services are "playing by the rules" while at the same time allowing them sufficient secrecy to fulfill their role? It seems feasible that technical mechanisms can contribute to solving this problem, and indeed a number of presentations addressed aspects of it. One might imagine that something analogous to the notion of zero-knowledge proofs might help demonstrate that intelligence agencies are following appropriate rules while not revealing details of those activities. Another possibility that was proposed is to make the amount of surveillance public in a verifiable fashion but without revealing the targets. Thus one might imagine that a specified limit be placed on the proportion of traffic available to intelligence services. The effect would be to force the agencies to be correspondingly selective in their choice of targets.

The crypto and security community should invest a substantial effort to make all layers

[1] This term is used to indicate Australia, Canada, UK, USA, and New Zealand.

of the internet and our devices more secure and to strengthen the level of privacy offered. This may create a natural barrier to mass surveillance and will also bring a more robust network infrastructure to a society that is increasingly reliant on it for critical services. Such a development may eventually increase the cost for targeted surveillance, but there is no indication that this would become prohibitive.

As is traditional for Dagstuhl, we started with a round table of quick introductions from the participants, including brief statements of what they hoped to get out of the workshop. We then had an open discussion on the goals of the workshop and of how best to organise the workshop to achieve these goals. It was decided to structure discussions into three strands:

- Principles
- Research directions
- Strategy

The outcomes of these discussions are detailed in a separate "Manifesto" document. The workshop was then structured into a number of plenary sessions alternating with breakouts into the three strands. The plenary sessions were made up of presentations from participants and feedback from the breakouts followed by discussion.

The problems addressed in this workshop are immensely challenging, and carry vast implications for society as a whole. It would not be reasonable to expect a small group of people – and a group not particularly representative of society as a whole – to produce solutions in the course of four days. Our goal was to gain some understanding of guiding principles and ways forward.

2 Table of Contents

Executive Summary

Talks

3 Talks

A total of 20 talks were given over course of the workshop, most of these taking around 30 minutes each. Abstracts for 18 of these talks are given below.

3.1 An introduction and brief overview of NSA and IC surveillance – from dragnets to drone strikes.

Jacob Appelbaum (The Tor Project, Cambridge US)

Surveillance is ultimately about power relationships. Various intelligence agencies wish to have total control through total surveillance. They attempt to sabotage cryptography in service of surveillance and censorship. Such power is used in concert with other agencies to perform actions ranging from harassment to political assassinations with drones.

3.2 How to Wiretap the Cloud without almost anybody noticing

Caspar Bowden (Independent privacy advocate, EU)

As Microsoft's Chief Privacy Adviser, I warned them about the effects of FISA 702 on the rest of the world's privacy in 2011. Shortly afterwards I was made redundant (and I did not know about PRISM, or that Microsoft was PRISM's first "corporate partner" since 2008).

My analysis was based on close scrutiny of open sources, and from Sep 2011 I tried to warn EU institutions, including the Commission (at the Cabinet level) and Data Protection Authorities. However no notice was taken. I contributed to a report to the European Parliament in Sep 2012 which laid out the precise legal mechanisms of FISA 702.

After the Snowden revelations the European Parliament phoned me up and said "Caspar . . . it's all true!" and asked me to write the official briefing Note for the EP inquiry. The analysis (if not all the Conclusions) was accepted in the official inquiry resolution and the findings of the Commission EU-US "Working Group' report.

A central conclusion is still under-reported. The actual definition in FISA 1801(e) of "foreign intelligence information" is conditioned on US nationality. A legal standard of "necessity" applies to Americans, but otherwise any information which relates to US foreign policy interests is caught. This structure appears to be unique. In the surveillance law of most other nations the distinction is based on international vs. domestic communications, but only a few other countries (AU, NZ, CA, DE) afford more rights to their own citizens.

This gives rise to extreme asymmetries in Cloud computing: for example, US citizens' data has equal protection to that of US residents' under European law, but the US recognizes no privacy rights in EU data reciprocally under US "national security" laws. The structure of FISA 702 amounts to a double-discrimination by nationality, which prima facie is incompatible with ECHR and ICCPR (but the US, UK, and Israel reject this interpretation).

There are also suspicious "FISA-shaped loopholes" buried deep in EU data protection laws, and more so even in the new proposed General Data Protection Regulation. Contrary to the mood music from the EU Commission that somehow the GDPR was the solution to post-Snowden privacy, several new bureaucratic mechanisms were invented to widen loopholes into floodgates. Similarly, the official EU DPAs were sanguine about Cloud computing before Snowden, in the face of clear warnings, but then were obliged to issue two "clarifications" on Cloud computing which have been some of the most richly amusing "Privacy theatre" on the stage. The last fifteen years of work by Data Protection institutions has crumbled into an abyss, and must now be reconstructed on foundations of computer science. In private, DPAs admit this.

3.3 Rigorous Survey on Privacy Attitudes Toward Privacy in Products

Jon Callas (Silent Circle and Blackphone, CH)

It is commonly asserted that privacy is a niche concern, that general consumers don't care enough about privacy to make it a significant factor in their decision to choose one product over another. But what is the reality, especially in the present world of surveillance? This presentation will show data collected from a statistically significant populations in the US and Germany, controlled for sex and broad population versus ICT-savvy people that suggests that whatever the anecdotes are, general consumers care about privacy.

3.4 On-line Privacy post-Snowden: what future? A European perspective

Joe Cannataci (University of Malta, MT, and University of Groningen, NL)

Like Charles Raab[2], I think that there is a good deal of work to be done on further conceptualising privacy in the context of security & open government, but I don't think that sorting out our conceptualisation of privacy will be key to doing something practical about privacy protection and surveillance in the short and mid-term.

In recent EU-supported research projects such as CONSENT[3], SMART[4], and RESPECT[5], the results of quantitative and qualitative research which involved thousands of EU citizens from across all 28 EU member states, suggest that:

- People care deeply about privacy even though their conceptualisation of it may be imprecise
- People care about privacy ... or say they do but will indulge in privacy-unfriendly behaviour especially if that behaviour is more convenient, easier, the path of least resistance etc. etc.

[2] Charles D. Raab, "Regulating surveillance: the importance of principles," in *Surveillance in Europe*, David Wright and Reinhard Kreissl, Routledge 2014; and Charles D> Raab, "Beyond the Privacy Paradigm: Implications for Regulating Surveillance," http://privacylaw.berkeleylawblogs.org/2013/05/24/charles-raab-beyond-the-privacy-paradigm-implications-for-regulating-surveillance/

[3] http://www.consent.law.muni.cz/

[4] http://smartsurveillance.eu/

[5] http://respectproject.eu/

- People care about surveillance, don't like pervasive surveillance . . . but their level of trust in the state varies depending on which state may live in;

When considering the prevailing set of realities and emerging trends it would seem that:

- on the one hand, in the world of big data, privacy invasion is often the business model;
- on the other hand, the very corporations which have grown huge on the back of advertising revenue derived from their ability to target individual customers, thanks to the profiling of their on-line activities and preferences, are ironically doing their utmost to retain or attract customers by introducing and leading the adoption of military-grade encryption in data transfer and data "at rest" whether in e-mail or other form of communication apps;

In this presentation, I use the MAPPING project[6] as a case study and an example of the latest European policy initiatives in cyberspace. MAPPING is a "Science in Society" project which is investigating practical ways forward at the points of intersection of internet governance, privacy protection and intellectual property rights. It is currently exploring the feasibility and desirability of a blended approach as the way forward by investigating:

- Technological solutions such as encryption;
- Overlay software solutions and eventually underlay architectural change as a method of improving user privacy and possibly by creating "parallel internets" or "parallel universes";
- Innovative legal instruments, especially in the sphere of international law, including a multi-lateral treaty which could serve as a new "magna charta for the internet"

When tackling the options listed above it is clear that a few home-truths need to be faced:

- Governments, organised crime and large corporations will continue to attempt surveillance against anybody/everybody;
- Governments will only "come to the table" if they have to, if they are made to do so;
- Technology, including cryptography, can be one of the ways to bring governments to the table;
- Personal data has become part of the business model for a multi-billion dollar per year industry;
- Re-thinking the business model is necessary but will encounter very stiff resistance from those with vested interests including their political allies;
- Having convenient (i.e. super-easy to use, low/non-cost) crypto is part of the answer;

A number of EU states do see eye-to-eye on the matter but they must co-exist in a space which, surveillance-wise, is currently dominated by the US, the UK and is assailed by at least ten (10) other nation states most of which are outside Europe. So what should Europe do? Build its own cyberspace under its own value-system and hope that others will eventually join it? Additionally, to complicate matters there's the spectre of cyberwar. Do you arm for it? Do you use it as an excuse? Or do you outlaw cyberwar? That is, do you declare cyberspace as a zone where no warfare shall be carried out in the spirit of, say, the START (nuclear) process or the Chemical Warfare treaty? The current stalemate in USA-EU relations over the matter of data protection and privacy law especially in the area of national security and mass surveillance, means that the European Commission's suggestion that the USA sign and ratify the Council of Europe's 1981 Data Protection Convention is unrealistic and insensitive to US domestic traditions and priorities. This suggests that new avenues need to be explored to try to resolve the stalemate and improve momentum on achieving a new consensus position. The MAPPING project is striving to explore such avenues in a number of ways. These

[6] http://www.mappingtheinternet.eu/

include the creation of a Working Group on Technical Solutions for cyberspace considering the feasibility and desirability of parallel internets where cryptography is not only welcome but indeed the norm. MAPPING has also created a Working Group on Legal Solutions which is investigating the feasibility and desirability of new legal instruments including that of a multilateral convention on surveillance in cyberspace. In doing so it raises the issue of which cyberspace? The current one or a cyberspace divided into a number of inter-connected but distinct networks subject to different jurisdictions. This extends the vision of the Internet as a "network of networks" to that of a "network of networks of networks".

The work in the MAPPING Project will be taken forward by meetings of its WP4 Technical Working Party on 12–13 November 2014 in Berlin and the WP4 Legal Working Party in Paris on 15–16 December 2015. Both Working Parties will then meet in Washington DC in the USA on 23–25 March 2015 with other meetings planned to be held in other locations including possibly Beijing. These Working Party meetings will prepare the ground for the MAPPING Annual General Assembly for stakeholders scheduled for 22–23 Sep 2015 and subsequent MAPPING General Assemblies for stakeholders scheduled for 2016, 2017 and 2018. The success of the European MAPPING project will depend on its continued ability to engage stakeholders world-wide and rise above the factors that have bogged down other internet governance and on-line privacy initiatives.

3.5 The technical and public policy ramifications of the Snowden revelations

George Danezis (University College London, GB)

The documents leaked to journalists by Edward Snowden have dominated the news for over a year, however the fragmented way in which they have been published makes it difficult to define their overarching narrative, and assess the overall impact of their substance. In this talk I will present a unified view of the pervasive monitoring operation of the NSA and GCHQ, spanning from different modes of access, to advanced analysis of the collected material. This unified approach will lead me directly to a number of technology public policy options for countries subject to such surveillance to protect their citizens and create incentives for a cyber-investigation regime that is more compatible with rule of law.

3.6 DP5: Privacy-preserving Presence Protocols

Ian Goldberg (University of Waterloo, CA)

Joint work of Borisov, Nikita; Danezis, George; Goldberg, Ian
Main reference N. Borisov, G. Danezis, I. Goldberg, "DP5: A Private Presence Service," Technical Report 2014-10, Centre for Applied Cryptographic Research, University of Waterloo, 2014.
URL http://cacr.uwaterloo.ca/techreports/2014/cacr2014-10.pdf

Users of social applications like to be notified when their friends are online. Typically, this is done by a central server keepingtrack of who is online and offline, as well as of the complete friendgraph of users. However, recent NSA revelations have shown that addressbook and buddy list information is routinely targetted for massinterception. Hence, some social service providers, such as activistorganizations, do not want to even possess this information about their users, lest it be taken or compelled from them. In this talk, we present DP5, a new suite of privacy-preserving presence protocols that allow people to determine when their friends are online (and to establish secure communications with them), without acentralized provider ever learning who is friends with whom. DP5 accomplishes this using an implementation of private information retrieval (PIR), which allows clients to retrieve information fromonline databases without revealing to the database operators what information is being requested.

3.7 We fix the Net!

Christian Grothoff (TU München, DE)

GCHQ, CSET and the NSA are colonizing the Internet, and the IETF is unwilling to commit to serious changes to the architecture to stop mass surveilance. The GNUnet project develops an alternative network architecture, initially to be deployed as an overlay network, to help civilization escape from PRISM.

3.8 Procurement – Privacy and Security

Marit Hansen (Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein – Kiel, DE)

Procurement procedures can be very influential concerning the choice for or against privacy and security features and guarantees. This talk discusses legal requirements from European data protection law as well as further advanced regional data protection legislation that demands that preference is given to products or systems that can prove compliance with law (State Data Protection Act Schleswig-Holstein from 2000) with a reference to certification procedure.

The data protection framework is not addressing security-relevant information, though. This became clear when journalists, on the basis of the Snowden revelations, reported on companies being active on behalf of the German government on the federal and on the state level that have connections with the NSA or are at least legally obliged to disclose privacy- or security-related information to government bodies (e.g. secret services). The reaction of the German federal government was a No-spy decree demanding a self-declaration of tenderers, and in fact they stopped the contract with some companies. This may be an interesting and potentially effective approach. However, it may be legally challenged as to be discriminating and not being in line with the European Procurement Directive.

3.9 The Tragedy of Privacy: Abstract and request for feedback

Amir Herzberg (Bar-Ilan University, IL)

Privacy is mostly viewed as a right of every *individual* to restrict collection, distribution and use of information related to him or herself. Most regulations, technologies and debate related to privacy, focus on this aspect, i.e., allowing individuals to control access to their private information, mainly by requiring *informed consent.* However, we argue that the informed-consent requirement may not suffice to protect the interests of individuals. Furthermore, we argue that there is insufficient attention to the *important role of privacy for society* as a whole; this implies a need for establishing additional privacy-protecting mechanisms.

We focus on the privacy threats due to the growing popularity of Big-Data Web-Services, which provide (usually free) services to huge populations, in return to permission to use user's information for commercial purposes such as advertising. We analyze potential applications of the big-data services, and argue that these corporations offer an increasingly-attractive deal for consumers, making the big-data services the ultimate resellers and providers of goods, services and information.

We argue that these advantages turn Big-Data Web-Services into a significant risk to global economy and society. The risk is due to their huge competitive advantage, and further differences compared to traditional businesses, mainly: (1) huge entrance barrier making this an exclusive club of few huge companies, (2) modest employee base and (3) mobility of assets and production facilities, allowing them to optimize location for tax-minimization and other considerations.

We argue that, in spite of concerns regarding personal privacy, it is perfectly *rational* for users to give their consent, even if they conceive dire implications from the establishment of huge collections of private information; we draw parallel to the well-known *tragedy of the commons*, and hence refer to this phenomena as the *tragedy of privacy.* This situation is aggravated due to cognitive processes that result in failure of individuals to properly evaluate the long-term implications of the sharing of private information, by themselves and by society at large, and due to the huge media-influence and branding of the Big-Data Web Services. The limited, arguably insufficient 'price' demanded by users for their information, is supported by experimental evidence, as well as by observations over existing practices.

We conclude that society should protect privacy by legislative measures, however, we offer a pessimistic forecast on the feasibility of adoption of effective measures that will restrict the collection and use of private information, properly protecting individuals and society. We note that the feasibility of adoption is further reduced by the cooperation between big-data services and governmental surveillance and intelligence agencies, which will further increase as gradual adoption of encryption technologies will reduce the value of eavesdropping. Another factor which works against adoption of privacy-protecting legislation is the growing use and dependency of politicians on big-data services, both processes further eroding privacy. One hope may be that governments may act to restrict privacy exposures to *foreign* corporations.

3.10 Back to the typewriters? – Rethinking informational self-determination in the era of mass state surveillance

Eleni Kosta (Tilburg University, NL)

The right to informational self-determination encompasses the right of individuals to determine who can use their data, for what purposes, under what conditions, and for how long. In private relationships, this is expressed primarily via individuals' consent to the processing of their personal data. Recent publications on the PRISM and TEMPORA surveillance programmes demonstrate that citizen data are being secretly – without their knowledge and consent – collected by the state via private companies, at a massive scale. This blanket and mass citizen surveillance seriously undermines individuals' informational self-determination and effective legal protection. The current checks and balances in consumer-business relationships are based on the assumption that government access to industry-processed data is an exception, which does not require regulation in the consumer-business realm. Now that the exception is effectively becoming the rule, the current legal framework of citizen protection presents a major gap. Therefore, consumer-business data protection requires rethinking. Building on the theories of informational self-determination and constitutionalisation of data protection, the proposed research will identify the required systemic revisions in the European data protection framework, in particular in the system of checks and balances to compensate for the loss of citizen control over state access to citizen data via private companies. For identifying the systemic revisions needed to compensate for this loss of control, rights and principles such as due process, transparency and accountability will be studied, as fundamental elements of the European legal tradition. The proposed research will contribute to transparency in business practices, to enhance legal certainty for users and companies. It will also be of great value to regulators and policy-makers, providing guidance on how the legal and regulatory framework can be adapted to offer effective legal protection when states access citizen data via the databases of private companies.

3.11 Search on Encrypted Data Can Be Practical (Use It!)

Hugo Krawczyk (IBM TJ Watson Research Center – Hawthorne, US)

Joint work of Cash, David; Faber, Sky; Jaeger, Joseph; Jarecki, Stanislaw; Jutla, Charanjit; Krawczyk, Hugo; Nguyen, Quan; Rosu, Marcel; Steiner, Michael

I will discuss some advances in practical solutions to the problem of searchableencryption in which a data owner outsources a database to an external server E (e.g., the cloud) in encrypted form. Later D can authorize clients to search thedata at E while hiding information about the database and queried values from E,and preventing clients from learning information they are not authorized for. We also consider the PIR-like requirement by which the data owner D needs to authorize queries while minimizing the information it learns about queried values. In all cases, searches at E are performed without ever decrypting dataor queries (in particular, E never gets the decryption keys). A solution developed by IBM Research and UC Irvine teams presents a majoradvance relative to prior work that focused on single-keyword search, singleclient, and implementations in small-size databases. This new work supportssearch via any boolean expression applied to sets of keywords associated withdocuments in the database as well as range queries. Our implementation of theproposed protocol has been tested on databases with billions of index entries (document-keyword pairs), e.g., a US-scale census database with 100 millionrecords each with 1,000 associated keywords and a very large collection ofcrawled web-pages that includes, among others, a full snapshot of the English Wikipedia. Recently, we expanded the system to support more costly queries including substring, wildcard and phrase searches. The availability of such technology enables privacy solutions that secure databy separating encrypted data from the keys use to decrypt it, and by applying fine-grain access control and delegation mechanisms at the data owner end. The data is secured against insider and outsider attacks on the outsourced database: the holder of such database cannot disclose the information even atgun point, it simply has no access to it. Applications range from conformance toprivacy regulations, secure and controlled sharing of information, and defenses against surveillance (even when the data is stored at foreign entities). This work is documented in the following papers:

- http://eprint.iacr.org/2013/169,
- http://eprint.iacr.org/2013/720,
- http://eprint.iacr.org/2014/853.

3.12 Dancing with Governments

Susan Landau (Department of Social Science and Policy Studies, Worcester Polytechnic Institute, US)

We have the tools and they have the problems, yet there's a mismatch. They don't take our solutions and we don't solve their problems. If the research community really wants to provide scientific advice to government, there are a number of steps to take, the first of which is to understand their problems – this is the actual problem, not the one they think they have. The second is to understand their equities, the third, to speak their language. This talk will discuss successfully dancing with governments, improving law, policy, and – sometimes – even privacy in the process.

3.13 Dual EC and what it taught us about vulnerabilities of the standardization ecosystem

Tanja Lange (TU Eindhoven, NL)

Joint work of S. Checkoway, R. Niederhagen, A. Everspaugh, M. Green, T. Lange, T. Ristenpart, D. J. Bernstein, J. Maskiewicz, H. Shacham, M. Fredrikson

Main reference S. Checkoway, R. Niederhagen, A. Everspaugh, M. Green, T. Lange, T. Ristenpart, D. J. Bernstein, J. Maskiewicz, H. Shacham, M. Fredrikson, "On the Practical Exploitability of Dual EC in TLS Implementations," in Proc. of the 23rd USENIX Security Symposium, pp. 319–335, USENIX Association, 2014.

URL https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/checkoway

URL https://projectbullrun.org/dual-ec

This talk describes how the back door in Dual EC works and how it can be exploited in TLS implementations. It also gives some historical background on how the standards including Dual EC came to live and when knowledge of the back door became public.

3.14 Security of Symmetric Encryption against Mass Surveillance

Kenneth G. Paterson (Royal Holloway University of London, GB)

Joint work of Bellare, Mihir; Paterson, Kenneth G.; Rogaway, Phillip
Main reference M. Bellare, K. Paterson, P. Rogaway, "Security of Symmetric Encryption against Mass Surveillance," in Proc. of the 34th Annual Cryptology Conf. (CRYPTO'14), LNCS, Vol. 8616, pp. 1–19, Springer, 2014; pre-print available as report 2014/438 from Cryptology ePrint Archive.
URL http://dx.doi.org/10.1007/978-3-662-44371-2_1
URL http://eprint.iacr.org/2014/438

Motivated by revelations concerning population-wide surveillance of encrypted communications, we formalize and investigate the resistance of symmetric encryption schemes to mass surveillance. The focus is on algorithm-substitution attacks (ASAs), where a subverted encryption algorithm replaces the real one. We assume that the goal of "big brother" is undetectable subversion, meaning that ciphertexts produced by the subverted encryption algorithm should reveal plaintexts to big brother yet be indistinguishable to users from those produced by the real encryption scheme. We formalize security notions to capture this goal and then offer both attacks and defenses. In the first category we show that successful (from the point of view of big brother) ASAs may be mounted on a large class of common symmetric encryption schemes. In the second category we show how to design symmetric encryption schemes that avoid such attacks and meet our notion of security. The lesson that emerges is the danger of choice: randomized, stateless schemes are subject to attack while deterministic, stateful ones are not.

3.15 Privacy as a Social Value and as a Security Value

Charles Raab (University of Edinburgh, UK)

Main reference Charles D. Raab, "Privacy as a Security Value,", pp. 39–58, in: Dag Wiese Schartum, Lee Bygrave and Anne Gunn Berge Bekken (eds.), "Jon Bing: En Hyllest / A Tribute", Oslo: Gyldendal, 2014.

- Privacy and security (national security, public safety) have no agreed singular meanings.
- Law and conventional wisdom: Privacy is only an individual right.
- But the social and public interest value of privacy is insufficiently recognized: it can be construed as a constitutive public good and as part of the public interest as well as being an individual right.
- Across and within societies, different people construe, and value, privacy differently, and privacy has to be seen contextually.
- Security is also a slippery term, and can refer to different levels of social scale: individual, neighbourhood, local community, a whole country or society, a region, the world. 'Safety' is a related concept (and in today's world, has become a pre-eminent value, along with 'security').
- These definitional and conceptual ambiguities and variations are not necessary a problem, except when – in legal, political, social, and medial parlance-it is said that (e.g.) 'privacy' conflicts with 'security' and must be 'balanced'.
- The concept of 'balance' is conceptually and empirically flawed; 'balancing' (individual) privacy and (national) security is a rhetorical and tendentious proposition.

- We need to think more imaginatively about the relationship between privacy and security, especially if we are to avoid security (almost always) trumping privacy in public policy, surveillance practice, and in popular parlance.
- It is helpful to consider that privacy and security have closer affinities than the 'versus' rhetoric allows. An important part of the value of privacy is that it affords a zone of security or safety. If so, the relationship between the two values is much more interesting and complex, and points to a need for a creative policy discourse.
- In addition, if privacy is an element of the public interest and a foundational principle of social relationships, for which there is considerable psychological and sociological support, then a relationship between privacy and security (or safety) as also public-interest values, becomes more complex and requires more subtle public-policy approaches.

3.16 End-to-end encrypted mail made easy for users

Mark D. Ryan (University of Birmingham, GB)

Main reference M. D. Ryan, "Enhanced certificate transparency and end-to-end encrypted mail," in Proc. of the 21st Annual Network and Distributed System Security Symposium (NDSS'14), 14 pages, The Internet Society, 2014.
Main reference http://www.internetsociety.org/doc/enhanced-certificate-transparency-and-end-end-encrypted-mail
URL https://www.cs.bham.ac.uk/~mdr/research/papers/pdf/14-ndss-cert.pdf

The certificate authority model for authenticating public keys of websites has been attacked in recent years, and several proposals havebeen made to reinforce it. We develop and extend *certificate transparency*, a proposal in this direction, so that it efficiently handles certificate revocation. We show how this extension can be usedto build a secure end-to-end email or messaging system using PKI with no requirement to trust certificate authorities, or to rely on complexpeer-to-peer key-signing arrangements such as PGP. This makesend-to-end encrypted mail possible, with apparently few additional usability issues compared to unencrypted mail (specifically, users donot need to understand or concern themselves with keys or certificates). Underlying these ideas is a new attacker model appropriate for cloud computing, which we call "malicious-but-cautious".

3.17 Data Obfuscation/Pollution: adapting TrackMeNot to counter surveillance

Vincent Toubiana

Joint work of Toubiana, Vincent; Howe, Daniel; Nissenbaum, Helen
URL http://cs.nyu.edu/trackmenot/

TrackMeNot is a browser extension designed to pollute the web search profile and web search history of users. Initial design of TrackMeNot considered search engines as the main adversaries. However, recent revelation about the NSA program XKeyScore highlights that surveillance can be triggered by specific search queries. This revelation raises the question "Could data pollution be used to make bulk collection inefficient?". Addressing this question implies to adapt the threat model to consider an adversary that use less accurate profiles. Furthermore, in order to adapt to this type of adversary it is necessary to find new sources of keywords like the list released by DHS in 2012. The open question is could data pollution have a positive outcome and could it be extended to other services than search.

3.18 The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions

Dan S. Wallach (Rice University, US)

Joint work of Zhu, Tao; Phipps, David; Pridgen, Adam; Crandall, Jedidiah R.; Dan S. Wallach
Main reference T. Zhu, D. Phipps, A. Pridgen, J.',R. Crandall, D. S. Wallach, "The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions," in Proc. of the 22th USENIX Security Symposium, pp. 227–240, USENIX Association, 2014; pre-print available from the author's webpage.
URL https://www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/zhu
URL http://www.cs.unm.edu/~crandall/usenix13.pdf

Weibo and other popular Chinese microblogging sites are well known for exercising internal censorship, to comply with Chinese government requirements. This research seeks to quantify the mechanisms of this censorship: how fast and how comprehensively posts are deleted. Our analysis considered 2.38 million posts gathered over roughly two months in 2012, with our attention focused on repeatedly visiting "sensitive" users. This gives us a view of censorship events within minutes of their occurrence, albeit at a cost of our data no longer representing a random sample of the general Weibo population. We also have a larger 470 million post sampling from Weibo's public timeline, taken over a longer time period, that is more representative of a random sample.

We found that deletions happen most heavily in the first hour after a post has been submitted. Focusing on original posts, not reposts/retweets, we observed that nearly 30% of the total deletion events occur within 5–30 minutes. Nearly 90% of the deletions happen within the first 24 hours. Leveraging our data, we also considered a variety of hypotheses about the mechanisms used by Weibo for censorship, such as the extent to which Weibo's censors use retrospective keyword-based censorship, and how repost/retweet popularity interacts with censorship. We also used natural language processing techniques to analyze which topics were more likely to be censored.

4 Acknowledgements

Our deepest thanks to Matt Blaze, who co-organized this Dagstuhl event with us, but who was unable to attend for reasons beyond his control. Many thanks to Johana Hamilton for making available to us her documentary film *1971*, which we were delighted to screen during our workshop prior to its theatrical release.

5 References

- "Necessary and Proportionate Principles." International Principles on the Application of Human Rights to Communications Surveillance. Final version, May 2014. Available from https://necessaryandproportionate.org/.
- Federal Trade Commission (USA). Privacy Online: A Report to Congress. June 1998. Available from the FTC website.
- Gary T. Marx. An Ethics for the New Surveillance. *The Information Society*, 14(3), pp. 171–186, 1998.

- Global Government Surveillance Reform. Joint from AOL, Apple, Dropbox, Facebook, Google, LinkedIn, Microsoft, Twitter, and Yahoo! – https://www.reformgovernmentsurveillance.com/
- Frank la Rue. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. Report to the United Nations General Assembly, Human Rights Council. A/HRC/23/40.

Participants

- Jacob Appelbaum
The Tor Project – Cambridge, US
- Daniel J. Bernstein
Univ. of Illinois – Chicago, US
- Caspar Bowden
GB
- Jon Callas
Silent Circle – San Jose, US
- Joseph Cannataci
University of Malta, MT & University of Groningen, The Netherlands
- George Danezis
University College London, GB
- Pooya Farshim
RHUL – London, GB
- Joan Feigenbaum
Yale University, US
- Ian Goldberg
University of Waterloo, CA
- Christian Grothoff
TU München, DE
- Marit Hansen
ULD SH – Kiel, DE
- Amir Herzberg
Bar-Ilan University – Ramat Gan, IL
- Eleni Kosta
Tilburg University, NL
- Hugo Krawczyk
IBM TJ Watson Res. Center – Hawthorne, US
- Susan Landau
Worcester Polytechnic Inst., US
- Tanja Lange
TU Eindhoven, NL
- Kevin S. McCurley
Google – San Jose, US
- David Naccache
ENS, Paris, FR
- Kenneth G. Paterson
Royal Holloway University of London, GB
- Bart Preneel
KU Leuven and iMinds, BE
- Charles D. Raab
University of Edinburgh, GB
- Phillip Rogaway
Univ. of California – Davis, US
- Mark D. Ryan
University of Birmingham, GB
- Peter Y. A. Ryan
University of Luxembourg, LU
- Haya Shulman
TU Darmstadt, DE
- Vanessa Teague
The University of Melbourne, AU
- Vincent Toubiana
CNIL – Paris, FR
- Michael Waidner
TU Darmstadt, DE
- Dan Wallach
Rice University, US

Report from Dagstuhl Seminar 14402

Resilience in Exascale Computing

Edited by

Hermann Härtig[1], Satoshi Matsuoka[2], Frank Mueller[3], and Alexander Reinefeld[4]

1 **TU Dresden, Germany,** haertig@os.inf.tu-dresden.de
2 **Tokyo Institute of Technology, Japan,** matsutitech@gmail.com
3 **North Carolina State University, USA,** mueller@cs.ncsu.edu
4 **Zuse Institute Berlin, Germany,** reinefeld@zib.de

Abstract

From September 28 to October 1, 2014, the Dagstuhl Seminar 14402 "Resilience in Exascale Computing" was held in Schloss Dagstuhl – Leibniz Center for Informatics. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available. Slides of the talks and abstracts are available online.

Seminar September 28 to October 1, 2014 – http://www.dagstuhl.de/14402
1998 ACM Subject Classification B.3.3 Performance Analysis and Design Aids, B.5.1 Design, C.1.2 Multiple Data Stream Architectures (Multiprocessors), D.1.3 Concurrent Programming
Keywords and phrases Exascale computing, resilience, fault tolerance, manycore computers, operating systems, micro kernels, work-load balancing, checkpointing
Digital Object Identifier 10.4230/DagRep.4.9.124

1 Executive Summary

Hermann Härtig
Satoshi Matsuoka
Frank Mueller
Alexander Reinefeld

Motivation

The upcoming transition[1] from petascale to exascale computers requires the development of radically new methods of computing. Massive parallelism, delivered by manycore processors and their assembly to systems beyond 10^7 processing units will open the way to extreme computing with more than 10^{18} floating point operations per second. The large number of functional components (computing cores, memory chips, network interfaces) will greatly increase the probability of partial failures. Already today, each of the four fastest supercomputers in the TOP500 list[2] comprises more than half a million CPU cores, and this tendency

[1] IDC top ten market prediction no. 2: *The Global Petascale/Exascale Race Will Keep Shifting the Market Toward Larger Systems*, IDC, March 2013.
[2] http://www.top500.org

towards massive parallelism is expected to accelerate in the future. In such large and complex systems, component failures are the norm rather than an exception. Applications must be able to handle dynamic reconfigurations during runtime and system software is needed to provide fault tolerance (FT) at a system level. For example, Jaguar reportedly experience 20 faults per hour in production mode[3], some of which could be mitigated while others could not.

To prevent valuable computation to be lost due to failures, checkpoint/restart (C/R) has become a requirement for long running jobs. However, current C/R mechanisms are insufficient, because the communication channels between the main memory and the parallel file system are far too slow to allow to save (and restore) a complete memory dump to disk farms. As an alternative, the memory of neighboring compute nodes may be used to keep partial checkpoints, but then erasure coding must be used to prevent against the loss of data in case of single node failures. To make things worse, precious communication bandwidth is needed for writing/reading checkpoints, which slows down the application. Techniques for data compression or application-specific checkpointing (with a reduced memory footprint) were proposed as a solution, but they only alleviate the problem by a certain extent.

We assume exascale hardware architectures to consist of a heterogeneous set of computational units (ranging from general-purpose CPUs to specialized units such as today's GPUs), memory chips (RAM, flash, phase-change memory), and various kinds of interconnects. The operating system and its load balancing mechanisms need to adapt to the hardware's properties as well as to workload characteristics. With the co-existence of legacy applications and new applications, it can be assumed that exascale systems must be capable of executing a broad range of parallel programming paradigms like MPI, OpenMP, PGAS, or MapReduce. These will not always and in every case require the functionality of a fully fledged operating system. We furthermore expect applications to become more complex and dynamic. Hence, developers cannot be expected to continuously handle load balancing and reliability. It is the operating system's task to find a sweet spot that on the one hand provides generic means for load management and checkpointing, while on the other hand allowing application developers full control over the performance-relevant functionality if required.

Objectives and Expected Results

The objective of this seminar is to bring together researchers and developers with a background on HPC system software (OS, network, storage, management tools) to discuss medium to long-term approaches towards resilience in exascale computers. Two concrete outcomes are (a) outlines for alternatives for resilience at extreme scale with trade-offs and dependencies on hardware/technology advances and (b) initiation of a standardization process for a resilience API. The latter is driven by current trends of resilience libraries to let users specify important data regions required for tolerating faults and for potential recovery. Berkeley Lab's BLCR, Livermore's SCR and Capello's FTI feature such region specification in their APIs, and so do may in-house application-specific solutions. A standardized resilience API would allow application programmers to be agnostic of future underlying resilience mechanisms and policies so that resilience libraries can be exchanged at will (and might even become inter-operable). The focus of solutions is on the practical system side and should reach

[3] A. Geist, "What is the Monster in the Closet?", August 2011, Invited Talk at Workshop on Architectures I: Exascale and Beyond: Gaps in Research, Gaps in our Thinking

beyond currently established solutions. Examples of areas of interest are:

- What is the "smallest denominator" that defines a resilience API? How can the standardization of a resilience API be realized?
- How can reactive FT schemes that respond to failures be enhanced to reduce system overhead, ensure progress in computation and sustain ever shorter MTBFs?
- How should low-energy and/or persistent memory be included on nodes for checkpointing (for example PCM) and used by applications and the OS?
- Can a significant number of faults be predicted with exact locations ahead of time so that proactive FT may provide complementary capabilities to move computation away from nodes that are about to fail?
- Can message logging, incremental checkpointing and similar techniques contribute to lower checkpointing overhead?
- Is redundant execution a viable alternative at exascale? How can partly redundant execution contribute to increased resilience in exascale algorithms?
- Can algorithm-based fault tolerance be generalized to entire classes of algorithms? Can results continuously be checked?
- What is the impact of silent data corruption (SDC) on HPC computing today? Which solvers can tolerate SDCs, which ones need to be enhanced (and how)?
- How do current/novel network architectures interact with the OS (e. g., how does migration interact with RDMA)?
- How can execution jitter be reduced or tolerated on exascale systems, particularly in the presence of failures?
- Can an interface be designed that allows the application to give "hints" to the OS in terms of execution steering for resilience handling? How does this approach interact with scalability mechanisms and policies, e. g., load balancing, and with programming models, e. g., to define fault handlers?
- Do distributed communication protocols offer better resilience? How do they support coordination between node-local and inter-node scheduling?
- Does "dark silicon" offer new opportunities for resilience?
- How can I/O on exascale be efficient and resilient (e. g., in situ analysis of simulation results)?

As a result of the seminar, we expect that this list of objectives will be refined, extended, and approaches to address each of these problems will be formulated, We anticipate that participants engage in increased coordination and collaboration within the currently (mostly) separate communities of HPC system software and application development.

Furthermore, the standardization process will be kicked off. One challenge is to find the most promising context for standardization. Current HPC-related standards (MPI, OpenMP, OpenACC) do not seem suitable since resilience cuts across concrete runtime environments and may also extend beyond HPC to Clouds and data centers involving industry participants from these area (in future standardization meetings beyond the scope of this meeting).

Overall, the objective of the workshop is to spark research and standardization activities in a coordinated manner that can pave the way for tomorrow's exascale computers to the benefit of the application developers. Thus we expect not only HPC system developers to benefit by the seminar but also the community of scientific computing at large, well beyond computer science. Due to the wide range of participants (researchers and industry practitioners from the U.S., Europe, and Asia), forthcoming research work may significantly help enhance FT properties of exascale systems, and technology transfer is likely to also reach general-purpose computing with many-core parallelism and server-style computing. Specifically, the work should set the seeds for increased collaborations between institutes in Europe and the U.S./Asia.

Relation to Previous Dagstuhl Seminars

Two of the proposers, Frank Mueller and Alexander Reinefeld, previously co-organized a Dagstuhl Seminar on *Fault Tolerance in High-Performance Computing and Grids* in 2009. It provided a forum for exchanging research ideas on FT in high-performance computing and grid computing community. In the meantime, the state-of-the-art greatly advanced and it became clear, that exascale computing will not be possible without adequate means for resilience. Hence, the new seminar will be more concrete in that the pressing problems of FT for exascale computing and standardization must be tackled and solved with the joint forces of system researchers and developers.

The proposed seminar also builds on the Dagstuhl Perspective Workshop 12212 *Co-Design of Systems and Applications for Exascale*, which also relates to the DFG-funded project FFMK (http://ffmk.tudos.org/, "A Fast and Fault-tolerant Microkernel-based System for Exascale Computing", DFG priority program 1648). Compared to the perspective workshop, our proposed seminar is much more focused on single, pressing topic of exascale computing, namely resilience.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

APIs, Checkpoint/Restart Systems and Resilience Benchmarks

3.1 Energy-Performance Tradeoffs in Multilevel Checkpoint Strategies

Leonardo Bautista-Gomez (Argonne National Laboratory, US)

Resilience in high-performance computing is all about protecting information. How to protect information while minimizing time, space and energy; is an open question. In this talk I will be presenting our recent work in multilevel checkpointing, lossy floating point compression, power monitoring and silent data corruption.

3.2 APIs, Architecture and Modeling for Extreme Scale Resilience

Kento Sato (Tokyo Institute of Technology, JP)

The computational power of high performance computing systems is growing exponentially, enabling finer grained scientific simulations. However, as the capability and component count of the systems increase, the overall failure rate increases accordingly. To make progress in spite of system failures, applications periodically write checkpoints to a reliable parallel file system so that the applications can restart from the last checkpoint. While simple, this conventional approach can impose huge overhead on application runtime for both checkpoint and restart operations at extreme-scale. In this presentation, to address the problem, first we introduce multi-level asynchronous checkpointing for fast checkpointing. Second, we present a fault tolerant messaging interface for fast and transparent recovery. Finally we explore new storage designs for scalable checkpoint/restart.

3.3 Portable Programming and Runtime Support for Application-Controlled Resilience

Andrew A. Chien (University of Chicago, US)

Application-aware techniques for resilience are promising approaches to efficient resilience in exascale systems. The Global View Resilience (GVR) system supports flexible, scalable, application-controlled resilience with a simple, portable abstraction – versioned, distributed arrays. Using a GVR prototype, we have evaluated the system's utility on both a number of mini-apps (miniMD, miniFE), and larger applications (a preconditioned conjugate gradient solver, OpenMC, ddcMD, and Chombo). Our results show that the programmer effort required (code change) to adopt version-based resilience is small (<1% code modified) and localized, requiring no software architecture changes. The application changes are also portable (machine-independent) and create a gentle-slope path to tolerating growing error rates in future systems. With the same applications, we evaluate the overhead of version-based resilience based on an early prototype GVR system, and find that low overheads can be achieved for all of them.

3.4 Open Discussion: Of Apples, Oranges and (Non-)reproducability

Frank Mueller (North Carolina State University, US)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.MuellerFrank.Slides.pdf

Current resilience work in HPC lacks a common API and common benchmarks. The objective of this discussion was to present and adapt an API that embraces most if not all resilience methods to date in a minimal set of routines, and to identify potential candidates for resilience benchmarks, execution scenarios with and without fault injection and metrics to report.

HPC Resilience methods and Beyond

3.5 MPI Fault Tolerance: The Good, The Bad, The Ugly

Martin Schulz (LLNL – Livermore, US)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.SchulzMartin.Slides.pdf

The MPI forum is currently investigating the inclusion of fault tolerance as a feature in the MPI specification. This issue has raised and continuous to raise some controversy about what MPI implementations can reasonably be expected to provide and what is useful for application developers. In this talk I will present the main proposals that are currently on the table, their advantages and the main concerns against them. The goal of this talk is to expand the discussion and to gather feedback that will help the MPI forum to come to a solution that is helpful for the larger HPC community.

3.6 Supporting the Development of Resilient Message Passing Applications

Christian Engelmann (Oak Ridge National Lab., US)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.EngelmannChristian.Slides.pdf

An emerging aspect of high-performance computing (HPC) hardware/software co-design is investigating performance under failure. The presented work extends the Extreme-scale Simulator (xSim), which was designed for evaluating the performance of message passing interface (MPI) applications on future HPC architectures, with fault-tolerant MPI extensions proposed by the MPI Fault Tolerance Working Group. xSim permits running MPI applications with millions of concurrent MPI ranks, while observing application performance in a simulated extreme-scale system using a lightweight parallel discrete event simulation. The newly added features offer user-level failure mitigation (ULFM) extensions at the simulated MPI layer to support algorithm-based fault tolerance (ABFT). The presented solution permits investigating performance under failure and failure handling of ABFT solutions. The newly enhanced xSim is the very first performance tool that supports ULFM and ABFT.

3.7 Fault Tolerance for Remote Memory Access Programming Models

Torsten Hoefler (ETH Zürich, CH)

Remote Memory Access (RMA) is an emerging mechanism for programming high-performance computers and datacenters. However, little work exists on resilience schemes for RMA-based applications and systems. In this paper we analyze fault tolerance for RMA and show that it is fundamentally different from resilience mechanisms targeting the message passing (MP) model. We design a model for reasoning about fault tolerance for RMA, addressing both flat and hierarchical hardware. We use this model to construct several highly-scalable mechanisms that provide efficient low- overhead in-memory checkpointing, transparent logging of remote memory accesses, and a scheme for transparent recovery of failed processes. Our protocols take into account diminishing amounts of memory per core, one of major features of future exascale machines. The implementation of our fault-tolerance scheme entails negligible additional overheads. Our reliability model shows that in-memory checkpointing and logging provide high resilience. This study enables highly-scalable resilience mechanisms for RMA and fills a research gap between fault tolerance and emerging RMA programming models.

3.8 Application level asynchronous checkpointing/restart: first experiences with GPI

Gerhard Wellein (Universität Erlangen-Nürnberg, DE)

Automatic check-point/restart is one potential way to address the resilience challenge of exascale computing. This talk presents first experiences of a prototypical application that is able to recover itself after detecting a failed node. It has been accomplished using GPI communication library in combination with checkpoint/restart technique. We investigate the potential of asynchronous checkpointing both to neighboring nodes and parallel file systems to reduce or even hide the costs of checkpointing.

Resilience Models

3.9 Operating System Support for Redundant Multithreading

Björn Döbel (TU Dresden, DE)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.DoebelBjoern.Slides.pdf

Implementing fault tolerance methods in software allows to protect commercial-off-the-shelf computer systems against the effects of transient and permanent hardware errors. In this talk I am going to present ROMAIN, an operating system service that provides transparent replication to binary-only applications on top of the L4 microkernel. I will describe how ROMAIN supports replication of multithreaded applications and replication of accesses to shared-memory channels between applications.

In the second part of the talk I am going to discuss how the experiences we gained while developing ROMAIN might benefit the HPC community and which of ROMAIN's assumptions need to be adjusted in this context.

3.10 Resilient gossip algorithms for online management of exascale clusters

Amnon Barak (The Hebrew University of Jerusalem, IL)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.BarakAmnon.Slides.pdf

Management of forthcoming exascale clusters requires frequent collection and sharing of runtime information about the health of the nodes, their resources and the running applications. We present a new paradigm for online management of scalable clusters, consisting of a large number of computing nodes (nodes) and a small number of servers (masters) that manage these nodes. We describe the details of gossip algorithms for sharing local information within subsets (colonies) of nodes and for sending selected global information to the masters, which hold recent information on all the nodes.

The presented algorithms are decentralized and resilient – they can work well when some nodes are down and without needing any recovery protocol. We give formal expressions for approximating the average age of the local information at each node and the average age of the collected information at a master. We then show that the results of these approximations closely match the results of simulations and measurements on a real cluster for different-size colonies.

The main outcome of this study is that a division of a large cluster to colonies can reduce the overall number of messages and the average load at the masters.

3.11 FFMK: Towards a fast and fault-tolerant micro-kernel-based Operating System

Hermann Härtig (TU Dresden, DE)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.HaertigHermann.Slides.pdf

FFMK's architecture is based on a combination of well-proven technologies: the L4 microkernel and L4-based virtualization, on-line management algorithms as used in the MosiX system, coding as used in RAID, and the XTreem-FS distributed file-system. I plan to explain the overall architecture and then start discussing the points in the architecture that require special attention for fault tolerance. My hope is, that – in interaction with the audience – we will arrive at a clearer idea where which types of fault-tolerance measures are needed or possible in view of requirements and fault models for exa-scale systems.

3.12 Open Discussion: A Holistic Model for Resilience

Hermann Härtig (TU Dresden, DE)

The following discussion suggested that a failure-model-based systematic analysis is needed rather than an ad-hoc discussion. It was suggested that the notion of *containment domains* makes sense as a starting point. Overall, a holistic view on fault tolerance techniques for exascale systems does yet exist.

Soft Errors

3.13 Memory Errors in Modern Systems

Vilas Sridharan (Advanced Micro Devices, Inc. – Boxborough, US)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.SridharanVilas.Slides.pdf

This talk presents fault and error data collected from several production systems in the field, and uses this data to project to node hardware reliability in an exascale timeframe.

3.14 Fault Tolerance for Iterative Linear Solvers

James J. Elliott (North Carolina State University, US)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.ElliottJames.Slides.pdf

Computer hardware trends may expose incorrect computation or storage to application codes. Silent data corruption (SDC) will likely be infrequent, yet one SDC suffices to make numerical algorithms like iterative linear solvers cease progress towards the correct answer. Initially, we focus our efforts on the resilience of the iterative linear solver GMRES to a single transient SDC. Our experiments show that when GMRES is used as the inner solver of an inner-outer iteration, it can "run through" SDC of almost any magnitude in the computationally intensive orthogonalization phase. That is, it gets the right answer using faulty data without any required roll back. Those SDCs, which it cannot run through, are caught by our detection scheme. We analyze our solvers in the presence of multiple faults, and discuss how fault rates and fault detection influences iterative solver selection.

3.15 Scalable Fault Tolerance at the Extreme Scale

Zizhong Chen (University of California – Riverside, US)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.ChenZizhong.Slides.pdf

Extreme scale supercomputers available before the end of this decade are expected to have 100 million to 1 billion computing cores. Due to the large number of components involved, extreme scale scientific applications must be protected from errors. When an error occurs, the affected application either continues or stops. If the application continues, we call it a fail-continue error. Otherwise, we call it a fail-stop error. In this talk, In this talk, I will discuss our recent work on scalable fault tolerance at the extreme scale. We have developed some highly efficient techniques for selected widely used scientific algorithms to tolerate both fail-continue and fail-stop errors according to their specific algorithmic characteristics. The algorithms we consider include direct methods for solving dense linear systems and eigenvalue problems, iterative methods for solving sparse linear systems and eigenvalue problems, and Newton's method for solving systems of non-linear equations and optimization problems. By leveraging the algorithmic characteristics of these algorithms, the proposed techniques can achieve much higher efficiency than the traditional general techniques (i. e., Triple Modular Redundancy for fail-continue errors and checkpoint for fail-stop errors) and therefore have potential to scale to exascale and beyond. A highly scalable checkpointing scheme is also developed for general applications.

3.16 Algorithms for coping with silent errors

Yves Robert (ENS – Lyon, FR & University of Tennessee, US)

Joint work of Guillaume Aupy, Anne Benoit, Aurélien Cavelan, Thomas Hérault, Saurabh K. Raina, Hongyang Sun, Frédéric Vivien, and Dounia Zaidouni
URL http://www.dagstuhl.de/mat/Files/14/14402/14402.RobertYves.Slides.pdf

Silent errors have become a major problem for large-scale distributed systems. Detection is hard, and correction is even harder. This talks presents generic algorithms to achieve both detection and correction of silent errors, by coupling verification mechanisms and checkpointing protocols. Application-specific techniques will also be investigated for sparse numerical linear algebra.

3.17 Assessing the impact of composite strategies for resilience

George Bosilca (University of Tennessee, US)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.BosilcaGeorge.Slides.pdf

With the advances in the theoretical and practical understanding of algorithmic traits enabling Algorithm Based Fault Tolerant (ABFT) approaches, a growing number of frequently used algorithms have been proven ABFT- capable. In the context of larger applications, these algorithms provide a temporal section of the execution when the data is protected by it's own intrinsic properties, and can be algorithmically recomputed without the need of checkpoints. However, while typical scientific applications spend a significant fraction of their execution time in library calls that can be ABFT- protected, they interleave sections that are difficult or even impossible to protect with ABFT. As a consequence, the only fault- tolerance approach that is currently used for these applications is checkpoint/restart. In this talk I will present a model to investigate the efficiency of a composite protocol, that alternates between ABFT and checkpoint/restart for effective protection of an iterative application composed of ABFT-aware and ABFT- unaware sections.

3.18 Leveraging PGAS Models for Hard and Soft Errors at Scale

Abhinav Vishnu (Pacific Northwest National Lab. – Richland, US)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.VishnuAbhinav.Slides.pdf

PGAS Models are finding increasing adoption in the community due to their productivity, asynchronous communication and high performance. In this talk, we will present research conducted at PNNL for leveraging PGAS programming models for hard faults and consider methods for soft error detection and correction using NWChem – a large scale computational chemistry application. A significant portion of the talk would be presenting the lessons learned and gaps which should be addressed in the resilience research.

3.19 Open Discussion: Soft Error

Satoshi Matsuoka (Tokyo Institute of Technology, JP)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.MatsuokaSatoshi.Slides.pdf

The so-called 'soft-errors' or undetected 'silent' errors are deemed to be one of the roadblocks as systems such as supercomputers and IDCs growing exponentially large, and the overall system error rate increasing proportionally. However, we often tend to forget that, unlike embedded, autonomous systems, for scientific computing there are always humans in the loop validating the results. In such a scenario, will soft errors matters as anticipated? We conduct qualitative analysis based on TSUBAME2.0's fault statics collected over a year period, to argue that soft errors may not be serious given the improvements in HW reliability as well as humans re-evaluating false positives due to soft errors, to the extent that the entire problem space could be dealt with traditional detection and recovery techniques for hard-stop failures.

Supporting Frameworks

3.20 Abstractions and mechanisms for proportional resilience

Mattan Erez (University of Texas – Austin, US)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.ErezMattan.Slides.pdf

Systems and applications have dynamic reliability characteristics and requirements. As a result, opportunities, and perhaps even a requirement, exist for co-tuning resilience across system and application layers. Many challenges must be addressed for co-tuning to be successful, especially in bridging the gaps between layers. In this talk I will discuss a few thoughts, examples, and questions related to mechanisms and abstractions (framed by Containment Domains) for addressing some of these challenges, including: how should hardware error and detector properties and models be presented to the application? How can an application specify non- traditional recovery and error tolerance? How can dynamic applications and load-balancing be modeled w.r.t. time and energy? What metrics should be optimized?

3.21 A Non-checkpoint/restart, Non-algorithm-specific Approach to Fault-tolerance

Dorian C. Arnold (University of New Mexico – Albuquerque, US)

URL http://www.dagstuhl.de/mat/Files/14/14402/14402.ArnoldDorian.Slides.pdf

Hierarchical or tree-based overlay networks (TBONs) are often used to execute data aggregation operations in a scalable, piecewise fashion. We present state compensation, a scalable failure recovery model for high-bandwidth, low-latency TBON computations. By leveraging inherently redundant state information found in many TBON computations, state compensation avoids explicit state replication (for example, process checkpoints and message logging) and incurs no overhead in the absence of failures. Further, when failures do occur, state compensation uses a weak data consistency model and localized protocols that allow processes to recover from failures independently and responsively. We describe the fundamental state compensation concepts and a prototype implementation integrated into the MRNet TBON infrastructure. Our experiments with this framework suggest that for TBONs supporting up to millions of application processes, state compensation can yield millisecond recovery latencies and inconsequential application perturbation.

3.22 Dynamic Resource Management and Scheduling for Fault Tolerance

Felix Wolf (GRS for Simulation Sciences – Aachen, DE)

To dynamically recover from node failure, a parallel job usually needs to replace the failed nodes. While the static allocation of spare nodes is technically simpler, pooling spare nodes across all jobs and allocating them dynamically is more efficient in terms of the number of required spare nodes but requires dynamic resource management. In this talk, we present an extension of the Torque/Maui batch systems to support dynamic node allocation for running parallel jobs and discuss how it can support fault-tolerant applications in the re-acquisition of failed nodes. As long as the parallel runtime system can continue running the application with replaced nodes, they can be obtained from the batch system through an API.

4 Conclusion

Through the lively engagement of all participants, the seminar was very successful and conducted in a professional, friendly and collegiate atmosphere supported by the kind and helpful staff at Schloss Dagstuhl. Lively discussions continued every day well beyond meeting times. The group meeting in its three-day format combined with the cozy confinement of Dagstuhl provided an umbrella for thoughtful discussion that conferences or workshops cannot provide. This helped create a community feeling that could become a building block for a concerted effort to coordinate future research activities, cooperate in outreach effort and maximize everyone's productivity and impact in fulling together each one's unique expertise for a combined effort to successfully solve the grand challenges of resilience in exascale HPC and beyond. During the meeting, follow-up action items with community-building character were identified, as detailed in the online discussion notes of the discussion sessions. They include (a) creation of a mailing list to coordinate activities and disseminate information on FT in high-performance computing and related areas, (b) development a standard set of benchmarks and an API for arbitrary resilience mechanisms, and (c) organization of follow-on meetings for the community. Within one month of the seminar, action item (a) have been realized and the benchmarks set (b) and a follow-up meeting (c) are in the planning stages.

Participants

- Dorian C. Arnold
University of New Mexico – Albuquerque, US
- Amnon Barak
The Hebrew University of Jerusalem, IL
- Leonardo Bautista-Gomez
Argonne National Laboratory, US
- George Bosilca
University of Tennessee, US
- Zizhong Chen
University of California – Riverside, US
- Andrew A. Chien
University of Chicago, US
- Nathan DeBardeleben
Los Alamos National Lab., US
- Björn Döbel
TU Dresden, DE
- James J. Elliott
North Carolina State Univ., US
- Christian Engelmann
Oak Ridge National Lab., US
- Mattan Erez
University of Texas – Austin, US
- Hermann Härtig
TU Dresden, DE
- Torsten Hoefler
ETH Zürich, CH
- Larry Kaplan
Cray Inc. – Seattle, US
- Dieter Kranzlmüller
LMU München, DE
- Matthias Lieber
TU Dresden, DE
- Naoya Maruyama
RIKEN – Kobe, JP
- Satoshi Matsuoka
Tokyo Institute of Technology, JP
- Frank Mueller
North Carolina State Univ., US
- Alexander Reinefeld
Konrad-Zuse-Zentrum – Berlin, DE
- Yves Robert
ENS – Lyon, FR
- Robert B. Ross
Argonne National Laboratory, US
- Kento Sato
Tokyo Institute of Technology, JP
- Thorsten Schütt
Konrad-Zuse-Zentrum – Berlin, DE
- Martin Schulz
LLNL – Livermore, US
- Vilas Sridharan
Advanced Micro Devices, Inc. – Boxborough, US
- Thomas Steinke
Konrad-Zuse-Zentrum – Berlin, DE
- Jeffrey Vetter
Oak Ridge National Lab., US
- Abhinav Vishnu
Pacific Northwest National Laboratory – Richland, US
- Gerhard Wellein
Univ. Erlangen-Nürnberg, DE
- Felix Wolf
GRS for Simulation Sciences – Aachen, DE