DAGSTUHL REPORTS

Volume 3, Issue 8, August 2013

 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/dagrep

Publication date
December, 2013

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

Aims and Scope
The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.
In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:
- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

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

Contact
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de

Digital Object Identifier: 10.4230/DagRep.3.8.i
www.dagstuhl.de/dagrep

Report from Dagstuhl Seminar 13321

Reinforcement Learning

Edited by

Peter Auer[1], Marcus Hutter[2], and Laurent Orseau[3]

1 **Montanuniversität Leoben, AT,** `auer@unileoben.ac.at`
2 **Australian National University – Canberra, AU,** `marcus.hutter@anu.edu.au`
3 **AgroParisTech – Paris, FR,** `laurent.orseau@agroparistech.fr`

Abstract

This Dagstuhl Seminar also stood as the 11th European Workshop on Reinforcement Learning (EWRL11). Reinforcement learning gains more and more attention each year, as can be seen at the various conferences (ECML, ICML, IJCAI, ...). EWRL, and in particular this Dagstuhl Seminar, aimed at gathering people interested in reinforcement learning from all around the globe. This unusual format for EWRL helped viewing the field and discussing topics differently.

Seminar 04.–09. August, 2013 – www.dagstuhl.de/13321

1998 ACM Subject Classification I.2.6 Learning, I.2.8 Problem Solving, Control Methods, and Search, I.2.9 Robotics, I.2.11 Distributed Artificial Intelligence, G.3 Probability and Statistics (Markov processes)

Keywords and phrases Machine Learning, Reinforcement Learning, Markov Decision Processes, Planning

Digital Object Identifier 10.4230/DagRep.3.8.1

1 Executive Summary

Peter Auer
Marcus Hutter
Laurent Orseau

Reinforcement Learning (RL) is becoming a very active field of machine learning, and this Dagstuhl Seminar aimed at helping researchers have a broad view of the current state of this field, exchange cross-topic ideas and present and discuss new trends in RL. It gathered 38 researchers together. Each day was more or less dedicated to one or a few topics, including in particular: The exploration/exploitation dilemma, function approximation and policy search, universal RL, partially observable Markov decision processes (POMDP), inverse RL and multi-objective RL.This year, by contrast to previous EWRL events, several small tutorials and overviews were presented. It appeared that researchers are nowadays interested in bringing RL to more general and more realistic settings, in particular by alleviating the Markovian assumption, for example so as to be applicable to robots and to a broader class of industrial applications.This trend is consistent with the observed growth of interest in policy search and universal RL. It may also explain why the traditional treatment of the exploration/exploitation dilemma received less attention than expected.

2 Table of Contents

3 Overview of Talks

3.1 Preference-based Evolutionary Direct Policy Search

Robert Busa-Fekete (Universität Marburg, DE)

We introduce a preference-based extension of *evolutionary direct policy search* (EDPS) as proposed by Heidrich-Meisner and Igel [2]. EDPS casts policy learning as a search problem in a parametric policy space, where the function to be optimized is a performance measure like expected total reward, andevolution strategies (ES) such as CMA-ES [1] are used as optimizers. Moreover, since the evaluation of a policy can only be done approximately, namely in terms of a finite number of *rollouts*, the authors make use of *racing algorithms* [3] to control this number in an adaptive manner. These algorithms return a sufficiently reliable ranking over the current set of policies (candidate solutions), which is then used by the ES for updating its parameters and population. A key idea of our approach is to extend EDPS by replacing the *value-based* racing algorithm with a *preference-based* one that operates on a suitable ordinal preference structure and only uses pairwise comparisons between sample rollouts of the policies.

References

1 N. Hansen and S. Kern. Evaluating the CMA evolution strategy on multimodal test functions. In *Parallel Problem Solving from Nature-PPSN VIII*, pages 282–291, 2004.
2 V. Heidrich-Meisner and C. Igel. Hoeffding and Bernstein races for selecting policies in evolutionary direct policy search. In *Proceedings of the 26th International Conference on Machine Learning*, pages 401–408, 2009.
3 O. Maron and A.W. Moore. Hoeffding races: accelerating model selection search for classification and function approximation. In *Advances in Neural Information Processing Systems*, pages 59–66, 1994.

3.2 Solving Simulator-Defined MDPs for Natural Resource Management

Thomas G. Dietterich (Oregon State University, US)

Joint work of Dietterich, Thomas G.; Taleghan Alkaee, Majid; Crowley, Mark
Main reference T. G. Dietterich, M. Alkaee Taleghan, M. Crowley, "PAC Optimal Planning for Invasive Species Management: Improved Exploration for Reinforcement Learning from Simulator-Defined MDPs," in Proc. of the AAAI Conference on Artificial Intelligence (AAAI'13), AAAI Press. 2013.
URL http://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6478
URL http://web.engr.oregonstate.edu/~tgd/publications/dietterich-taleghan-crowley-pac-optimal-planning-for-invasive-species-management-etc-aaai2013.pdf

Natural resource management problems, such as forestry, fisheries, and water resources, can be formulated as Markov decisionprocesses. However, solving them is difficult for two reasons. First, the dynamics of the system are typically available only in the form of a complex and expensive simulator. This means that MDP planning algorithms are needed that minimize the number of calls to the simulator. Second, the systems are spatial. A natural way to formulate the MDP is to divide up the region into cells, where each cell is modeled with a small number of state variables. Actions typically operate at the level of individual cells, but the spatial dynamics couple the states of spatially-adjacent cells. The resulting state and action spaces of these MDPs are immense. We have been working on two natural resource MDPs. The first involves the spread of tamarisk in river networks. A native of the Middle East, tamarisk has become an invasive plant in the dryland rivers and streams of the western US. Given a tamarisk invasion, a land manager must decide how and where to fight the invasion (e.g., eradicate tamarisk plants? plant native plants? upstream? downstream?). Although approximate or heuristic solutions to this problem would beuseful, our collaborating economists tell us that our policy recommendations will carry more weight if they are provably optimal with high probability. A large problem instance involves 4.7×10^6 states with 2187 actions in each state. On a modern 64-bit machine, the action-value function for this problem can fit into main memory. However, computing the full transition function to sufficient accuracy to support standardvalue iteration requires on the order of 3×10^{20} simulator calls. The second problem concerns the management of wildfire in Eastern Oregon. In this region, prior to European settlement, the native ponderosa pine forests were adapted to frequent, low-intensity fires. These fires allow the ponderosa pine trees (which are well-adapted to survive fire) to grow very tall while preventing the accumulation of fuel at ground level. These trees provide habitat for many animal species and are also very valuable for timber. However, beginning in the early 1900s, all fires were suppressed in this landscape, which has led to the build up of huge amounts of fuel. The result has been large, catastrophic fires that kill even the ponderosa trees and that are exceptionally expensive to control. The goal of fire management is to return the landscape to a state where frequent, low-intensity fires are again the normal behavior. There are two concrete fire management problems: Let Burn (decide which fires to suppress) and Fuel Treatment (decide in which cells to perform fuel reduction treatments). Note that in these problems, the system begins in an unusual, non-equilibrium state, and the goal is to return the system to a desired steady state distribution. Hence, these problems are not problems of reinforcement learning, but rather problems of MDP planning for a specific start state. Many of the assumptions in RL papers, such as ergocity of all policies, are not appropriate for this setting. Note also that it is highly desirable to produce a concrete policy (as opposed to just producing near-optimal behavior viareceding horizon control). A concrete policy can be inspected by stakeholders to identify missing constraints, state variables, and components of the reward function. To solve these problems, we are exploring two lines of research. For tamarisk, we have been building on recent work in PAC-RL algorithms(e.g., MBIE, UCRL, UCRL2, FRTDP, OP) to develop PAC-MDP planning algorithms. We are pursuing two innovations. First, we have developed an exploration heuristic based on an upper bound on the discounted state occupancy probability. Second, we are developing tighter confidence intervals in order to terminate the search earlier. These are based on combining Good-Turing estimates of missing mass (i.e., for unseen outcomes) with sequential confidence intervals for multinomial distributions. These reduce the degree to which we must rely on the union bound, and hence give us tighter convergence. For the LetBurn wildfire problem, we are exploring approximate policy iteration methods. For Fuel Treatment, we are extending Crowley's Equilibrium Policy Gradient methods. These define a local policy function that stochastically chooses the action for cell i based on the actions already chosen for the cells in the surrounding neighborhood. A Gibbs-sampling-style MCMC method repeatedly samples from these local policies until a global equilibrium is reached. This equilibrium defines the global policy. At equilibrium, gradient estimates can be computed and applied to improve the policy.

3.3 ABC and Cover Tree Reinforcement Learning

Christos Dimitrakakis (EPFL – Lausanne, CH)

Joint work of Dimitrakakis, Christos; Tziortziotis, Nikolaos
Main reference C. Dimitrakakis, N. Tziortziotis, "ABC Reinforcement Learning," arXiv:1303.6977v4 [stat.ML], 2013; to appear in Proc. of ICML'13.
URL http://arxiv.org/abs/1303.6977v4

In this talk, I presented our recent results on methods for Bayesian reinforcement learning using Thompson sampling, but differing significantly on their prior. The first, Approximate Bayesian Computation Reinforcement Learning (ABC-RL), employs an arbitrary prior over a set of simulators and is most suitable in cases where an uncertain simulation model is available. The second,Cover Tree Bayesian Reinforcement Learning (CTB-RL), performs closed-form online Bayesian inference on a cover tree and is suitable for arbitrary reinforcement learning problems, when little is known about the environment and fast inference is essential. ABC-RL introduces a simple, general framework for likelihood-free Bayesian reinforcement learning, through Approximate Bayesian Computation (ABC). The advantage is that we only require a prior distribution ona class of simulators. This is useful when a probabilistic model of the underlying process is too complex to formulate, but where detailed simulation models are available. ABC-RL allows the use of any Bayesian reinforcement learning technique in this case. It can be seen as an extension of simulation methods to both planning and inference. We experimentally demonstrate the potential of this approach in a comparison with LSPI. Finally, we introduce a theorem showing that ABC is sound, in the sense that the KL divergence between the incomputable true posterior and the ABC approximation is bounded by an appropriate choice of statistics. CTBRL proposes an online tree-based Bayesian approach for reinforcement learning. For inference, we employ a generalised context tree model. This defines a distribution on multivariate Gaussian piecewise-linear models, which can be updated in closed form. The tree structure itself is constructed using the cover tree method, which remains efficient in high dimensional spaces. We combine the model with Thompson sampling and approximate dynamic programming to obtain effective exploration policies in unknown environments. The flexibility and computational simplicity of the model render it suitable for many reinforcement learning problems in continuous state spaces. We demonstrate this in an experimental comparison with least squares policy iteration.

References

1 N. Tziortziotis, C. Dimitrakakis and K. BlekasCover Tree Bayesian Reinforcement Learning. arXiv:1305.1809

3.4 Some thoughts on Transfer Learning in Reinforcement Learning: on States and Representation

Lutz Frommberger (Universität Bremen, DE)

Main reference L. Frommberger, "Qualitative Spatial Abstraction in Reinforcement Learning," Cognitive Technologies Series, ISBN 978-3-642-16590-0, Springer, 2010.
URL http://dx.doi.org/10.1007/978-3-642-16590-0

The term "transfer learning" is a fairly sophisticated term for something thatcan be considered a core component of any learning effort of a human or animal: to base the solution to a new problem on experience and learning success of prior learning tasks. This is something that a learning organism does implicitly from birth on: no task is ever isolated, but embedded in a common surrounding or history. In contrary to this lifelong learning type setting, transfer learning in RL[5] assumes two different MDPs $\mathcal{M}$ and $\mathcal{M}'$ that have something "in common". This commonality is mostlikely given in a task mapping function that maps states and actions from $\mathcal{M}$ to $\mathcal{M}'$ as a basis for reusing learned policies. Task mappings can be given by human supervisors or learned, but mostly there is some instance telling the learning agent what to do to benefit from its experience. In very common words: Here is task $\mathcal{M}$, there is task$\mathcal{M}'$, and this is how you can bridge between them. This is a fairly narrow view on information reuse, and more organic and autonomous variants of knowledge transfer are desirable. Knowledge transfer, may it be in-task (i.e., generalization) or cross-task, exploits similarity between tasks. By task mapping functions, information on similarity is brought into the learning process from outside. This also holds for approaches that do not require an explicit state mapping [2, 4], where relations or agent spaces, e.g., are defined a-priori. What is mostly lacking so far is the agent's ability to recognize similarities on its own and/or seamlessly benefit from prior experiences as an integral part of the new learning effort. An intelligent learning agent should easily notice if certain parts of the current task are identical or similar to an earlier learning task, for example, general movement skills that remain constant over many specialized learning tasks. In prior work, I proposed generalization approaches such as task space tilecoding [1] that allow to reuse knowledge of the actual learning task if certain state variables are identical. This works if structural information is made part of the state space and does not require a mapping function. However, it needs a-priori knowledge of which state variables are critical for action selection in a structural way. Recent approaches foster thehope that such knowledge can be retrieved by the agent itself: e.g.,[3] allows for identification of state variables that have a generally high impact on action selection over one or several tasks. But even if we can identify and exploit certain state variables that encode structural information and have this generalizing impact, these features must at least exist. If they do not exist in the state representations, such approaches fail. For example, if the distance to the next obstacle in front of an agent is this critical value, it does not help if the agent's position and the position of the obstacle are in the state representation, as this implicitly hides the critical information. Thus, again, the question of state space design becomes evident. How can we ensure that relevant information is encoded on the level of features? Or how can we exploit information that is only implicitly given in the state representation? Answering these questions will be necessary to take the next step into autonomous knowledge reuse for RL agents.

References

1 Lutz Frommberger. Task space tile coding: In-task and cross-task generalization in reinforcement learning. In *9th European Workshop on Reinforcement Learning (EWRL9)*, Athens, Greece, September 2011.

2 George D. Konidaris and Andrew G. Barto. Autonomous shaping: Knowledge transfer in reinforcement learning. In *Proceedings of the Twenty Third International Conference on Machine Learning (ICML 2006)*, pp. 489–49, Pittsburgh, PA, June 2006.

3 Matthijs Snel and Shimon Whiteson. Multi-task reinforcement learning: shaping and feature selection. In *Recent Advances in Reinforcement Learning*, pp. 237–248. Springer, 2012.

4 Prasad Tadepalli, Robert Givan, and Kurt Driessens. Relational reinforcement learning: An overview. In *Proc. of the ICML-2004 Workshop on Relational Reinforcement Learning*, pp. 1–9, 2004.

5 Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *The Journal of Machine Learning Research*, 10:1633–1685,2009.

3.5 Actor-Critic Algorithms for Risk-Sensitive MDPs

Mohammad Ghavamzadeh (INRIA Nord Europe – Lille, FR)

In many sequential decision-making problems we may want to manage risk by minimizing some measure of variability in rewards in addition to maximizing a standard criterion. Variance related risk measures are among the most common risk-sensitive criteria in finance and operations research. However, optimizing many such criteria is known to be a hard problem. In this paper, we consider both discounted and average reward Markov decision processes. For each formulation, we first define a measure of variability for a policy, which in turn gives us a set of risk-sensitive criteria to optimize. For each of these criteria, we derive a formula for computing its gradient. We then devise actor-critic algorithms for estimating the gradient and updating the policy parameters in the ascent direction. We establish the convergence of our algorithms to locally risk-sensitive optimal policies. Finally, we demonstrate the usefulness of our algorithms in a traffic signal control application.

3.6 Statistical Learning Theory in Reinforcement Learning and Approximate Dynamic Programming

Mohammad Ghavamzadeh (INRIA Nord Europe – Lille, FR)

Approximate dynamic programming (ADP) and reinforcement learning (RL) algorithms are used to solve sequential decision-making tasks where the environment (i.e., the dynamics and the rewards) is not completely known and/or the size of thestate and action spaces is too large. In these scenarios, the convergence and performance guarantees of the standard DP algorithms are no longer valid, and adifferent theoretical analysis have to be developed. Statistical learning theory (SLT) has been a fundamental theoretical tool to explain the interaction between the process generating the samples and the hypothesis space used by learning algorithms, and shown when and how-well classification and regression problems can be solved. In recent years, SLT tools have been used to study the performance of batch versions of RL and ADP algorithms with the objective of deriving finite-sample bounds on the performance loss (w.r.t. the optimal policy) of the policy learned by these methods. Such an objective requires to effectively combine SLT tools with the ADP algorithms, and to show how the error is propagated through the iterations of these iterative algorithms.

3.7 Universal Reinforcement Learning

Marcus Hutter (Australian National University, AU)

There is great interest in understanding and constructing generally intelligent systems approaching and ultimately exceeding human intelligence. Universal AI is such a mathematical theory of machinesuper-intelligence. More precisely, AIXI is an elegant parameter-free theory of an optimal reinforcement learning agent embedded in an arbitrary unknown environment that possesses essentially all aspects of rational intelligence. The theory reduces all conceptual AI problems to pure computational questions. After a brief discussion of its philosophical, mathematical, and computational ingredients, I will give a formal definition and measure of intelligence, which is maximized by AIXI. AIXI can be viewed as the most powerful Bayes-optimal sequential decision maker, for which I will present general optimality results. This also motivates some variations such as knowledge-seeking and optimistic agents, and feature reinforcement learning. Finally I present some recent approximations, implementations, and applications of this modern top-down approach to AI.

References

1 M. Hutter. *Universal Artificial Intelligence.* Springer, Berlin, 2005.
2 J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver. *A Monte Carlo AIXI approximation. Journal of Artificial Intelligence Research*, 40:95–142, 2011.
3 S. Legg. *Machine Super Intelligence. PhD thesis*, IDSIA, Lugano, Switzerland, 2008.
4 M. Hutter. *One decade of universal artificial intelligence.* In *Theoretical Foundations of Artificial General Intelligence*, pages 67–88. Atlantis Press, 2012.
5 T. Lattimore. *Theory of General Reinforcement Learning. PhD thesis*, Research School of Computer Science, Australian National University, 2014.

3.8 Temporal Abstraction by Sparsifying Proximity Statistics

Rico Jonschkowski (TU Berlin, DE)

Joint work of Jonschkowski, Rico; Toussaint, Marc;

Automatic discovery of temporal abstractions is a key problem in hierarchical reinforcement learning. In previous work, such abstractions were found through the analysis of a set of experienced or demonstrated trajectories [2]. We propose that, from such trajectory data, we may learn temporally marginalized transition probabilities—which we call proximity statistics—that model possible transitions on larger time scales rather than learning 1-step transition probabilities. Viewing the proximity statistics as state values allows the agent to generate greedy policies from them. Making the statistics sparse and combining proximity estimates by *proximity propagation* can substantially accelerate planning compared to value iteration while keeping the size of the statistics manageable. The concept of proximity statistics and its sparsification approach is inspired from recent work in transit-node routing in large road networks [1]. We show that sparsification of these proximity statistics implies an approach to the discovery of temporal abstractions. Options defined by subgoals are shown to be a special case of the sparsification of proximity statistics. We demonstrate the approach and compare various sparsification schemes in a stochastic grid world.

References

1 Holger Bast; Setfan Funke; Domagoj Matijevic; Peter Sanders; Dominik Schultes, *In transit to constant time shortest-path queries in road networks*, Workshop on Algorithm Engineering and Experiments, 46–59, 2007.
2 Martin Stolle; Doina Precup, *Learning options in reinforcement learning*, Proc. of the 5th Int'l Symp. on Abstraction, Reformulation and Approximation, pp. 212–223, 2002.

3.9 State Representation Learning in Robotics

Rico Jonschkowski (TU Berlin, DE)

Joint work of Jonschkowski, Rico; Brock, Oliver
Main reference R. Jonschkowski, O. Brock, "Learning Task-Specific State Representations by Maximizing Slowness and Predictability," in Proc. of the 6th Int'l Workshop on Evolutionary and Reinforcement Learning for Autonomous Robot Systems (ERLARS'13), 2013.
URL http://www.robotics.tu-berlin.de/fileadmin/fg170/Publikationen_pdf/Jonschkowski-13-ERLARS-final.pdf

The success of reinforcement learning in robotic tasks is highly dependent onthe state representation – a mapping from high dimensional sensory observations of the robot to states that can be used for reinforcement learning. Currently, this representation is defined by human engineers – thereby solving an essential part of the robotic learning task. However, this approach does not scale because it restricts the robot to tasks for which representations have been predefined by humans. For robots that are able to learn new tasks, we need state representation learning. We sketch how this problem can be approached by iteratively learning a hierarchy of task-specific state representations following a curriculum. We then focus on a single step in this iterative procedure: learning a state representation that allows the robot to solve asingle task. To find this representation, we optimize two characteristics of good state representations: predictability and slowness. We implement these characteristics in a neural network and show that this approach can find good state representations from visual input in simulated robotic tasks.

3.10 Reinforcement Learning with Heterogeneous Policy Representations

Petar Kormushev (Istituto Italiano di Tecnologia – Genova, IT)

Joint work of Kormushev, Petar; Caldwell, Darwin G.
Main reference P. Kormushev, D. G. Caldwell, "Reinforcement Learning with Heterogeneous Policy Representations," 2013.
URL http://ewrl.files.wordpress.com/2013/06/ewrl11_submission_20.pdf

We propose a novel reinforcement learning approach for direct policy search that can simultaneously: (i) determine the most suitable policy representation for a given task; and (ii) optimize the policy parameters of this representation in order to maximize the reward and thus achieve the task. The approach assumes that there is a heterogeneous set of policy representations available to choose from. A naïve approach to solving this problem would be to take the available policy representations one by one, run a separate RL optimization process (i.e. conduct trials and evaluate the return) for each once, and at the very end pick the representation that achieved the highest reward. Such an approach, while theoretically possible, would not be efficient enough in practice. Instead, our proposed approach is to conduct one single RL optimization process while interleaving simultaneously all available policy representations. This can be achieved by leveraging our previous work in the area of RL based on Particle Filtering (RLPF).

3.11 Theoretical Analysis of Planning with Options

Timothy Mann (Technion – Haifa, IL)

Joint work of Mann, Timothy; Mannor, Shie

We introduce theoretical analysis suggesting how planning can benefit from using options. Experimental results have shown that options often induce faster convergence [3, 4], and previous theoretical analysis has shown that options are well-behaved in dynamic programming[2, 3]. We introduced a generalization of the Fitted Value Iteration (FVI) algorithm [1] that incorporates samples generated by options. Our analysis reveals that when the given set of options contains the primitive actions, our generalized algorithm converges approximately as fast as FVI with only primitive actions. When only temporally extended actions are used for planning convergence can be significantly faster than planning with only primitives, but this method may converge toward a suboptimal policy. We also developed precise conditions where our generalized FVI algorithm converges faster with a combination of primitive and temporally extended actions than with only primitive actions. These conditions turn out to depend critically on whether the iterates produced by FVI underestimate the optimal value function. Our analysis of FVI suggests that options can play an important role in planning by inducing fast convergence.

References

1 Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.
2 Doina Precup, Richard S Sutton, and Satinder Singh. Theoretical results on reinforcement learning with temporallyabstract options. Machine Learning: ECML-98, Springer, 382–393, 1998.
3 David Silver and Kamil Ciosek. Compositional Planning Using Optimal Option Models. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, 2012.
4 Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112 (1):181–211, August 1999.

3.12 Learning Skill Templates for Parameterized Tasks

Jan Hendrik Metzen (Universität Bremen, DE)

Joint work of Metzen, Jan Hendrik; Fabisch, Alexander

We consider the problem of learning skills for a parameterized reinforcement learning problem class. That is, we assume that a task is defined by a task parameter vector and, likewise, a skill is considered as a parameterized policy. We propose skill templates, which allow to generalize skills that have been learned using reinforcement learning to similar tasks. In contrast to the recently proposed parameterized skills[1], skill templates also provide a measure of uncertainty for this generalization, which is useful for subsequent adaptation of the skill by means of reinforcement learning. In order to infer a generalized mapping from task parameter space to policy parameter space and an estimate of its uncertainty, we use Gaussian process regression [3]. We represent skills by dynamical movement primitives [2] and evaluate the approach on a simulated Mitsubishi PA10 arm, where learning a single skill corresponds to throwing a ball to a fixed target position while learning the skill template requires to generalize to new target positions. We show that learning skill templates requires only a small amount of training data and improves learning in the target task considerably.

References

1 B. C. da Silva, G. Konidaris, and A. G. Barto. *Learning Parameterized Skills*, 29th International Conference on Machine Learning, 2012.
2 A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal. *Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors*. Neural Computation, 25:2, 328–373, 2013.
3 C. E. Rasmussen, C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

3.13 Online learning in Markov decision processes

Gergely Neu (Budapest University of Technology & Economics, HU)

Joint work of Neu, Gergely; Szepesvari, Csaba; Zimin, Alexander; Gyorgy, Andras; Dick, Travis
Main reference A. Zimin, G. Neu, "Online learning in Markov decision processes by Relative Entropy Policy Search," to appear in Advances in Neural Information Processing 26.

We study the problem of online learning in finite episodic Markov decision processes (MDPs) where the loss function is allowed to change between episodes. The natural performance measure in this learning problem is the regret defined as the difference between the total loss of the best stationary policy and the total loss suffered by the learner. We assume that the learner is given access to a finite action space $\mathcal{A}$ and the state space $\mathcal{X}$ has a layered structure with L layers, so that state transitions are only possible between consecutive layers. We propose several learning algorithms based on applying the well-known Mirror Descent algorithm to the problem described above. For deriving our first method, we observe that Mirror Descent can be regarded as a variant of the recently proposed Relative Entropy Policy Search (REPS) algorithm of [1]. Our corresponding algorithm is called Online REPS or O-REPS. Second, we show how to approximately solve the projection operations required by Mirror Descent without taking advantage of the connections to REPS. Finally, we propose a learning method based on using a modified version of the algorithm of[2] to implement the Continuous Exponential Weights algorithm for the online MDP problem. For these last two techniques, we provide rigorous complexity analyses. More importantly, we show that all of the above algorithms satisfy regret bounds of $O(\sqrt{L\,|\mathcal{X}|\,|\mathcal{A}|\,T\log(|\mathcal{X}|\,|\mathcal{A}|\,/L)})$ in the bandit setting and $O(L\sqrt{T\log(|\mathcal{X}|\,|\mathcal{A}|\,/L)})$ in thefull information setting (both after T episodes). These guarantees largely improve previously known results under much milder assumptions and cannot be significantly improved under general assumptions.

References

1 Peters, J., Mülling, K., and Altün, Y. (2010). Relative entropy policy search. In *Proc. of the 24th AAAI Conf. on Artificial Intelligence (AAAI 2010)*, pp. 1607–1612.
2 Narayanan, H. and Rakhlin, A. (2010). Random walk approach to regret minimization. In *Advances in Neural Information Processing Systems 23*, pp. 1777–1785.

3.14 Hierarchical Learning of Motor Skills with Information-Theoretic Policy Search

Gerhard Neumann (TU Darmstadt – Darmstadt, DE)

Joint work of Neumann, Gerhard; Daniel, Christian; Kupcsik, Andras; Deisenroth, Marc; Peters, Jan
URL G. Neumann, C. Daniel, A. Kupcsik, M. Deisenroth, J. Peters, "Hierarchical Learning of Motor Skills with Information-Theoretic Policy Search," 2013.
URL http://ewrl.files.wordpress.com/2013/06/ewrl11_submission_1.pdf

The key idea behind information-theoretic policy search is to bound the 'distance' between the new and old trajectory distribution, where the relative entropy is used as 'distance measure'. The relative entropy bound exhibits many beneficial properties, such as a smooth and fast learning process and a closed-form solution for the resulting policy. We summarize our work on information theoretic policy search for motor skill learning where we put particular focus on extending the original algorithm to learn several options for a motor task, select an option for the current situation, adapt the option to the situation and sequence options to solve an overall task. Finally, we illustrate the performance of our algorithm with experiments on real robots.

3.15 Multi-Objective Reinforcement Learning

Ann Nowe (Free University of Brussels, BE)

Joint work of Nowe, Ann; Van Moffaert, Kristof; Drugan, M. Madalina
Main reference K. Van Moffaert, M. M. Drugan, A. Nowé, "Hypervolume-based Multi-Objective Reinforcement Learning," in Proc. of the 7th Int'l Conf. on Evolutionary Multi-Criterion Optimization (EMO'13), LNCS, Vol. 7811, pp. 352–366, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-37140-0_28

We focus on extending reinforcement learning algorithms to multi-objective problems, where the value functions are not assumed to be convex. In these cases, the environment – either single-state or multi-state – provides the agent multiple feedback signals upon performing an action. These signals can be independent, complementary or conflicting. Hence, multi-objective reinforcement learning (MORL) is the process of learning policies that optimize multiple criteria simultaneously. In our talk, we briefly describe our extensions to multi-armed bandits and reinforcement learning algorithms to make them applicable in multi-objective environments. In general, we highlight two main streams in MORL, i.e. either the scalarization or the direct Pareto approach. The simplest method to solve a multi-objective reinforcement learning problem is to use a scalarization function to reduce the dimensionality of the objective space to a single-objective problem. Examples are the linear scalarization function and the non-linear Chebyshev scalarization function. In our talk, we highlight that scalarization functions can be easily applied in general but their expressive power depends heavily on on the fact whether a linear or non-linear transformation to a single dimension is performed [2]. Additionally, they suffer from additional parameters that heavily bias the search process. Without scalarization functions, the problem remains truly multi-objective and$Q$-values have to be learnt for each objective individually and therefore a state-action is mapped to a Q-vector. However, a problem arises in the boots trapping process as multiple actions be can considered equally good in terms of the partial order Pareto dominance relation. Therefore, we extend the RL boots trapping principle to propagating sets of Pareto dominating Q-vectors in multi-objective environments. In [1], we propose to store the average immediate reward and the Pareto dominating future discounted reward vector separately. Hence, these two entities can converge separately but can also easily be combined with a vector-sum operator when the actual Q-vectors are requested. Subsequently, the separation is also a crucial aspect to determine the actual action sequence to follow a converged policy in the Pareto set.

References

1 K. Van Moffaert, M. M. Drugan, A. Nowé , *Multi-Objective Reinforcement Learning using Sets of Pareto Dominating Policies*,Conference on Multiple Criteria Decision Making, 2013.

2 M. M. Drugan, A. Nowé , *Designing Multi-Objective Multi-Armed Bandits: An Analysis*, in Proc of International Joint Conference of Neural Networks (IJCNN 2013).

3.16 Bayesian Reinforcement Learning + Exploration

Tor Lattimore (Australian National University – Canberra, AU)

Joint work of Lattimore, Tor; Hutter, Marcus

A reinforcement learning policy π interacts sequentially with an environment μ. In each time-step the policy π takes action $a \in \mathcal{A}$ before receiving observation $o \in \mathcal{O}$ and reward $r \in \mathcal{R}$. The goal of an agent/policy is to maximise some version of the (expected/discounted) cumulative reward. Since we are interested in the reinforcement learning problem we will assume that thetrue environment μ is unknown, but resides in some known set $\mathcal{M}$. The objective is to construct a single policy that performs well in some sense for all/most $\mu \in \mathcal{M}$. This challenge has been tackled for many specific $\mathcal{M}$, including bandits and factored/partially observable/regular MDPs, but comparatively few researchers have considered more general history-based environments. Here we consider arbitrary countable $\mathcal{M}$ and construct a principled Bayesian inspired algorithm that competes with the optimal policy in Cesaro average.

3.17 Knowledge-Seeking Agents

Laurent Orseau (AgroParisTech – Paris, FR)

Joint work of Orseau, Laurent; Lattimore, Tor; Hutter, Marcus
Main reference L. Orseau, T. Lattimore, M. Hutter, "Universal Knowledge-Seeking Agents for Stochastic Environments," in Proc. of the 24th Int'l Conf. on Algorithmic Learning Theory (ALT'13), LNCS, Vol. 8139, pp. 146–160, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-40935-6_12
URL http://www.hutter1.net/publ/ksaprob.pdf

Observing that the optimal Bayesian rational agent AIXI does not explore its environment entirely led us to give a seek a definition of an optimal Bayesian that does so in an optimal way. We recently defined such a knowledge-seeking agent, KL-KSA, designed for countable hypothesis classes of stochastic environments. Although this agent works for arbitrary countable classes and priors, we focus on the especially interesting case where all stochastic computable environments are considered and the prior is based on Solomonoff's universal prior. Among other properties, we show that KL-KSA learns the true environment in the sense that it learns to predict the consequences of actions it does not take. We show that it does not consider noise to be information and avoids taking actions leading to inescapable traps. We also present a variety of toy experiments demonstrating that KL-KSA behaves according to expectation.

3.18 Toward a more realistic framework for general reinforcement learning

Laurent Orseau (AgroParisTech – Paris, FR)

Joint work of Orseau, Laurent; Ring, Mark
Main reference L. Orseau, M. Ring, "Space-Time Embedded Intelligence," in Proc. of the 5th Int'l Conf. on Artificial General Intelligence (AGI'12), LNCS, Vol. 7716, pp. 209–218, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-35506-6_22

The traditional agent framework, commonly used in reinforcement learning (RL)and elsewhere, is particularly convenient to formally deal with agents interacting with their environment. However, this framework has a number of issues that are usually of minor importance but can become severe when dealing with general RL and artificial general intelligence, where one studies agents that are optimally rational, or can merely have a human-level intelligence. As a simple example, an intelligent robot that is controlled by rewards and punishments through a remote control should by all means try to get hold of this remote control, in order to give itself as many rewards as possible. In a series of paper [1, 2, 4, 3], we studied various consequences of integrating the agent more and more in its environment, leading to a new definition of artificial general intelligence where the agent is fully embedded in the world, to the point where it is even computed by it [3].

References

1 Ring, M., & Orseau, L. (2011). *Delusion, Survival, and Intelligent Agents.* Artificial General Intelligence (AGI) (pp. 11–20). Berlin, Heidelberg: Springer.
2 Orseau, L., & Ring, M. (2011). *Self-Modification and Mortality in Artificial Agents.* Artificial General Intelligence (AGI) (pp. 1–10). Springer.
3 Orseau, L., & Ring, M. (2012). *Space-Time Embedded Intelligence.* Artificial General Intelligence (pp. 209–218). Oxford, UK: Springer Berlin Heidelberg.
4 Orseau, L., & Ring, M. (2012). *Memory Issues of Intelligent Agents.* Artificial General Intelligence (pp. 219–231). Oxford, UK: Springer Berlin Heidelberg.

3.19 Colored MDPs, Restless Bandits, and Continuous State Reinforcement Learning

Ronald Ortner (Montan-Universität Leoben, AT)

Joint work of Ortner, Ronald; Ryabko, Daniil; Auer, Peter; Munos, Rémi
Main reference R. Ortner, D. Ryabko, P. Auer, R. Munos, "Regret Bounds for Restless Markov Bandits," in Proc. of the 23rd Int'l Conf. on Algorithmic Learning Theory (ALT'12), LNCS, Vol . 7568, pp. 214–228, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-34106-9_19

We introduce the notion of colored MDPs that allows to add structural information to ordinary MDPs. Thus, state-action pairs are assigned the same color when they exhibit similar rewards and transition probabilities. This extra information can be exploited by an adaptation of the UCRL algorithm, leading to regret bounds that depend on the number of colors instead of the size of the state-action space. As applications, we are able to derive regret bounds for the restless bandit problem as well as for continuous state reinforcement learning.

3.20 Reinforcement Learning using Kernel-Based Stochastic Factorization

Joëlle Pineau (McGill University – Montreal, CA)

Joint work of Barreto, André M.S.; Precup, D.; Pineau, Joëlle
Main reference A. M. S. Barreto, D. Precup, J. Pineau, "Reinforcement Learning using Kernel-Based Stochastic Factorization," in Proc. of the 25th Annual Conf. on Neural Information Processing Systems (NIPS'11), pp. 720–728, 2011.
URL http://books.nips.cc/papers/files/nips24/NIPS2011_0496.pdf
URL http://www.cs.mcgill.ca/~jpineau/files/barreto-nips11.pdf

Recent years have witnessed the emergence of several reinforcement-learning techniques that make it possible to learn a decision policy from a batch of sample transitions. Among them, kernel-based reinforcement learning (KBRL)stands out for two reasons. First, unlike other approximation schemes, KBRL always converges to a unique solution. Second, KBRL is consistent in the statistical sense, meaning that adding more data improves the quality of the resulting policy and eventually leads to optimal performance. Despite its nice theoretical properties, KBRL has not been widely adopted by the reinforcement learning community. One possible explanation for this is that the size of the KBRL approximator grows with the number of sample transitions, which makes the approach impractical for large problems. In this work, we introduce a novel algorithm to improve the scalability of KBRL. We use a special decomposition of a transition matrix, called stochastic factorization, which allows us to fix the size of the approximator while at the same time incorporating all the information contained in the data. We apply this technique to compress the size of KBRL-derived models to a fixed dimension. This approach is not only advantageous because of the model-size reduction; it also allows a better bias-variance trade-off, by incorporating more samples inthe model estimate. The resulting algorithm, kernel-based stochastic factorization (KBSF), is much faster than KBRL, yet still converges to a unique solution. We derive a theoretical bound on the distance between KBRL's solutionand KBSF's solution. We show that it is also possible to construct the KBSF solution in a fully incremental way, thus freeing the space complexity of the approach from its dependence on the number of sample transitions. The incremental version of KBSF (iKBSF) is able to process an arbitrary amount of data, which results in a model-based reinforcement learning algorithm that canbe used to solve large continuous MDPs in on-line regimes. We present experiments on a variety of challenging RL domains, including thedouble and triple pole-balancing tasks, the Helicopter domain, the penthatlon event featured in the Reinforcement Learning Competition 2013, and a model of epileptic rat brains in which the goal is to learn a neurostimulation policy to suppress the occurrence of seizures.

3.21 A POMDP Tutorial

Joëlle Pineau (McGill University – Montreal, CA)

URL http://www.cs.mcgill.ca/~jpineau/talks/jpineau-dagstuhl13.pdf

This talk presented key concepts, algorithms, theory and empirical results pertaining to learning and planning in Partially Observable Markov Decision Processes (POMDPs).

3.22 Methods for Bellman Error Basis Function construction

Doina Precup (McGill University – Montreal, CA)

Function approximation is crucial for obtaining good results in large reinforcement learning tasks, but the problem of devising a good function approximator is difficult and often solved in practice by hand-crafting the“right” set of features. In the last decade, a considerable amount of effort has been devoted to methods that can construct value function approximators automatically from data. Among these methods, Bellman error basis functionc onstruction (BEBF) are appealing due to their theoretical guarantees and good empirical performance in difficult tasks. In this talk, we discuss on-going developments of methods for BEBF construction based on random projections (Fard, Grinberg, Pineau and Precup, NIPS 2013) and orthogonal matching pursuit (Farahmand and Precup, NIPS 2012).

3.23 Continual Learning

Mark B. Ring (Anaheim Hills, US)

Joint work of Ring, Mark B.; Schaul, Tom; Schmidhuber, Jürgen
Main reference M. B. Ring, T. Schaul, “The organization of behavior into temporal and spatial neighborhoods,” in Proc. of the 2012 IEEE Int’l Conf. on Development and Learning and Epigenetic Robotics (ICDL-EPIROB’12), pp. 1–6, IEEE, 2012.
URL http://dx.doi.org/10.1109/DevLrn.2012.6400883

A continual-learning agent is one that begins with relatively little knowledge and few skills but then incrementally and continually builds up new skills and new knowledge based on what it has already learned. But what should such an agent do when it inevitably runs out of resources? One possible solution is to prune away less useful skills and knowledge, which is difficult if these are closely connected to each other in a network of complex dependencies. The approach I advocate in this talk is to give the agent at the outset all the computational resources it will ever have, such that continual learning becomes the process of continually reallocating those fixed resources. I will describe how an agent's policy can be broken into many pieces and spread out among many computational units that compete to represent different parts of the agent's policy space. These units can then be arranged across a lower-dimensional manifold according to those similarities, which results in many advantages for the agent. Among these advantages are improved robustness, dimensionality reduction, and an organization that encourages intelligent reallocation of resources when learning new skills.

3.24 Multi-objective Reinforcement Learning

Manuela Ruiz-Montiel (University of Malaga, ES)

Joint work of Ruiz-Montiel, Manuela; Mandow, Lawrence; Pérez-de-la-Cruz, José-Luis
Main reference M. Riuz-Montiel, L. Mandow, J. L. Pérez-de-la-Cruz, "PQ-Learning: Aprendizaje por Refuerzo Multiobjetivo," in Proc. of the XV Conference of the Spanish Association for Artificial Intelligence (CAEPIA'13), pp. 139–148, 2013.
URL http://www.congresocedi.es/images/site/actas/ActasCAEPIA.pdf

In this talk we present PQ-learning, a new Reinforcement Learning (RL)algorithm that determines the rational behaviours of an agent in multi-objective domains. Most RL techniques focus on environments with scalar rewards. However, many real scenarios are best formulated in multi-objective terms: rewards are vectors and each component stands for an objective to maximize. In scalar RL, the environment is formalized as a Markov Decision Problem, defined by a set S of states, a set A of actions, a function $P_{sa}(s')$ (the transition probabilities) and a function $R_{sa}(s')$ (the obtained scalar rewards). The problem is to determine a*policy* $\pi : S \rightarrow A$ that maximizes the *discounted accumulated reward* $R_t = \Sigma_{k=0}^{\infty} \gamma^k r_{t+k+1}$.E.g., Q-learning [1] is an algorithm that learns such policy. It learns the scalar values $Q(s,a) : S \times A \rightarrow \mathbb{R}$, that represent the expected accumulated reward when following a given policy after taking a in s. The selected action a in each state is given by the expression $argmax_a Q(s,a)$. In the multi-objective case the rewards are vectors $\overrightarrow{r} \in \mathbb{R}^n$, so different accumulated rewards cannot be totally ordered; $\overrightarrow{v}$ dominates $\overrightarrow{w}$ when $\exists i : v_i > w_i \wedge \nexists j : v_j < w_j$. Given a set of vectors, those that are not dominated by any other vector are said to lie in the *Pareto front*. We seek the set of policies that yield non-dominated accumulated reward vectors. The literature on multi-objective RL (MORL) is relatively scarce (see Vamplew et al. [2]). Most methods use preferences (lexicographic ordering or scalarization) allowing a total ordering of the value vectors, and approximate the front by running a scalar RL method several times with different preferences. When dealing with non-convex fronts, only a subset of the solutions is approximated. Some multi-objective dynamic programming (MODP) methods calculate all the policies at once, assuming a perfect knowledge of $P_{sa}(s')$ and $R_{sa}(s')$. We deal with the problem of efficiently approximating all the optimal policies at once, without sacrificing solutions nor assuming a perfect knowledge of the model. As far as we know, our algorithm is the first to bring these featurestogether. As we aim to learn a set of policies at once, Q-learning is apromising starting point, since the policy used to interact with theenvironment is not the same that is learned. At each step, Q-learning shifts the previous estimated Q-value towards its new estimation: $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma max_{a'} Q(s',a'))$. In PQ-learning, Q-values are sets of vectors, so the max operator is replaced by $ND(\bigcup_{a'} Q(s',a')))$, where ND calculates the Pareto front. A naive approach to perform the involved set addition is a pairwise summation (imported from MODP methods), but it leads to an uncontrolled growth of the sets and the algorithm becomes impractical, as it sums vectors that correspond to different action sequences. The results of these mixed sums are useless when learning deterministic policies, because two sequences cannot be followed at once. We propose a controlled set addition that only sums those pairs of vectors that correspond to useful action sequences. This is done by associating each vector $\overrightarrow{q}$ with two data structures with information about the vectors that (1) have been updated by $\overrightarrow{q}$ and (2) have contributed to its value. In this talk we describe in detail the application ofPQ-learning to a simple example, and the results that the algorithm yields when applied to two problems of a benchmark [2]. It approximates all the policies in the true Pareto front, as opposed to the naive approach, that produces huge fronts with useless values that dramatically slow down the process.[1]

References

1. C.J. Watkins, *Learning From DelayedRewards*. PhD Thesis, University of Cambridge, 1989.
2. P. Vamplew et al., *Empirical EvaluationMethods For Multiobjective Reinforcement Learning*, in Machine Learning84(1-2) pp. 51-80, 2011.

3.25 Recent Advances in Symbolic Dynamic Programming for Hybrid MDPs and POMDPs

Scott Sanner (NICTA – Canberra, AU)

Joint work of Sanner, Scott; Zamani, Zahra

Many real-world decision-theoretic planning problems are naturallymodeled using mixed discrete and continuous state, action, and observation spaces, yet little work has provided *exact* methodsfor performing exact dynamic programming backups in such problems. This overview talk will survey a number of recent developments in the exact and approximate solution of mixed discrete and continuous (hybrid) MDPs and POMDPs via the technique of symbolic dynamic programming (SDP) as covered in recent work by the authors [1, 2, 3, 4].

References

1. S. Sanner, K. V. Delgado, and L. Nunes de Barros. Symbolic dynamic programming for discrete and continuous state MDPs. In *In Proc. of the 27th Conf. on Uncertainty in Artificial Intelligence (UAI-11)*, Barcelona, Spain, 2011.
2. Z. Zamani, S. Sanner, K. V. Delgado, and L. Nunes de Barros. Robust optimization for hybrid mdps with state-dependent noise. In *Proc. of the 23rd International Joint Conf. on Artificial Intelligence (IJCAI-13)*, Beijing, China, 2013.

[1] This work is partially funded by: grant TIN2009-14179 (SpanishGovernment, Plan Nacional de I+D+i) and Universidad de Málaga, Campusde Excelencia Internacional Andalucía Tech. Manuela Ruiz-Montiel is funded by the Spanish Ministry of Education through the National F.P.U. Program.

3 Z. Zamani, S. Sanner, and C. Fang. Symbolic dynamic programming for continuous state and action mdps. In *In Proc. of the 26th AAAI Conf. on Artificial Intelligence (AAAI-12)*, Toronto, Canada, 2012.
4 Z. Zamani, S. Sanner, P. Poupart, and K. Kersting. Symbolic dynamic programming for continuous state and observation pomdps. In *In Proc. of the 26th Annual Conf. on Advances in Neural Information Processing Systems (NIPS-12)*, Lake Tahoe, Nevada, 2012.

3.26 Deterministic Policy Gradients

David Silver (University College – London, GB)

Joint work of David Silver

In this talk we consider deterministic policy gradient algorithms for reinforcement learning with continuous actions. The deterministic policy gradient has a particularly appealing form: it is the expected gradient of the action-value function. This simple form means that the deterministic policy gradient can be estimated much more efficiently than the usual stochastic policy gradient. To ensure adequate exploration, we introduce an off-policy actor-critic algorithm that learns a deterministic target policy from an exploratory behaviour policy. We demonstrate that deterministic policy gradient algorithms can significantly outperform their stochastic counterparts in high-dimensional action spaces.

3.27 Sequentially Interacting Markov Chain Monte Carlo BasedPolicy Iteration

Orhan Sönmez (Boğaziçi University – Istanbul, TR)

Joint work of Sönmez, Orhan; Cemgil, A. Taylan

In this ongoing research, we introduce a policy iteration method where policies are evaluated using sequentially interacting Markov chain Monte Carlo (SIMCMC) [1] for planning in discrete time continuous state space Markov decision processes (MDPs).In order to do so, we utilize the expectation-maximization algorithm derived for solving MDPs [2] and employ a SIMCMC sampling scheme in its intractable expectation step. Fortunately, the maximization step has a closed form solution due to Markov properties. Meanwhile, we approximate the policy as a function over the continuous state space using Gaussian processes [3]. Hence, in the maximization step, we simply select the state-action pairs of the trajectories sampled by SIMCMC as the support of the Gaussian process approximator. We are aware that SIMCMC methods are not the best choice with respect to sample efficiency compared to sequential Monte Carlo samplers (SMCS)[4]. However, they are more appropriate for online settings due to their estimation at any time property which SMCSs lack. As a future work, we are investigating different approaches to develop an online reinforcement learning algorithm based on SIMCMC policy evaluation. As a model based approach, the dynamics of the model would be approximated with Gaussian processes [5].

References

1 Brockwell, A., Del Moral, P., Doucet A., Sequentially Interacting Markov Chain Monte Carlo Methods. emphThe Annals of Statistics, 38(6):3870–3411, December 2010.
2 Toussaint, M., Storkey, A., Probabilistic Inference for Solving Discrete and Continuous State Markov Decision Processes. In *Proc. of the 23rd Int'l Conf. on MachineLearning*, pp. 945–952, 2006.
3 Deisenroth, M., Rasmussen C.E., Peters, J., Gaussian Process Dynamic Programming. *Neurocomputing*, 72(7-9):1508–1524,March 2009.
4 Del Moral, P., Doucet A., Sequential Monte Carlo Samplers. *Journal of the Royal Statistical Society – Series B: StatisticalMethodology*, 68(3):411–436, June 2006.
5 Deisenroth, M., Rasmussen C.E.PILCO: A Model-Based and Data-Efficient Approach to Policy Search. In *Proc. of the 28th Int'l Conf. on Machine Learning*, pp. 465–472, 2006.

3.28 Exploration versus Exploitation in Reinforcement Learning

Peter Sunehag (Australian National University, AU)

Main reference T. Lattimore, M. Hutter, P. Sunehag, "The Sample-Complexity of General Reinforcement Learning," arXiv:1308.4828v1 [cs.LG]; and in Proc. of the 30th Int'l Conf. on Machine Learning (ICML'13), JMLR W&CP 28(3):28–36, 2013.
URL http://arxiv.org/abs/1308.4828v1
URL http://jmlr.org/proceedings/papers/v28/lattimore13.pdf

My presentation was a tutorial overview of the exploration vs exploitation dilemma in reinforcement learning. I began in the multi-armed bandit setting and went through Markov Decision Processes to the general reinforcement learning setting that has only recently been studied. The talk discussed the various strategies for dealing with the dilemma like optimismin a frequentist setting or posterior sampling in the Bayesian setting, as well as the performance measures like sample complexity in discounted reinforcement learning or regret bounds for undiscounted the setting. Itwas concluded that sample complexity bounds can be proven in much more general settings than regret bounds. Regret bounds need some sort of recoverability guarantees while unfortunately sample complexity says less about how much reward the agent will achieve. The speaker's recommendation is to try to achieve optimal sample complexity but only within the class of rational agents described by an axiomatic system developed from classical rational choice decision theory.

3.29 The Quest for the Ultimate TD(λ)

Richard S. Sutton (University of Alberta – Edmonton, CA)

Joint work of Maei, H. R.; Sutton, R. S.
Main reference H. R. Maei, R. S. Sutton, "GQ(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces," in Proc. of the 3rd Conf. on Artificial General Intelligence (AGI'10), Advances in Intelligent Systems Research Series, 6 pp., 2010.
URL http://dx.doi.org/10.2991/agi.2010.22
URL http://webdocs.cs.ualberta.ca/ sutton/papers/maei-sutton-10.pdf

TD(λ) is a computationally simple model-free algorithm for learning to predict long term consequences. It has been used to learn value functions, to form multi-scale models of the world, and to compile planning into policies for immediate action. It is a natural core algorithm for artificial intelligence based on reinforcement learning. Before realizing its full potential, however, TD(λ) needs to be generalized in several ways: to off-policy learning, as has already been partially done, to maximally general parameterization, as has also been partially done, and to off-policy eligibility traces,which was previously thought impossible but now perhaps we can see how this too can be done. In all these ways we see a glimmer of a perfected and complete algorithm—something inspired by TD(λ), with all its positive computational and algorithmic features, and yet more general, flexible, and powerful. Seeking this perfected algorithm is the quest for the ultimate TD (λ); this talk is a status report on the goals for it and the prospects for achieving them.

References

1 H. R. Maei, R. S. Sutton. GQ(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. *In Proceedings of the Third Conference on Artificial General Intelligence*, Lugano, Switzerland, 2010.

3.30 Relations between Reinforcement Learning, Visual Input, Perception and Action

Martijn van Otterlo (Radboud University Nijmegen, NL)

Typical reinforcement learning (RL) algorithms learn from *traces* of state, action, state, action, . . . sequences, in order to optimize action selection for each state (wrt. a reward criterium). The field of RL has come up with many algorithms based on abstraction and generalization[7]. In my view general RL amounts to feedback-based, *interactive experimentation* with particular abstraction levels forstates, actions and tasks. However, despite all previous efforts direct couplings of RL with complex visual input (e.g. raw images) are still rare. In roughly the recent decade, RL has been combined with so-called *relational* knowledge representation for states and languages[5, 6]. Also, many forms of *decision-theoretic planning*, using abstract or relational version of Bellman equations, can employ powerful knowledge representation schemes[4]. An interesting development is that also in the computervision community, people wish to employ similar relational generalization overvisual input, due to advances in (probabilistic) logical learning (e.g. see our recent work on the interpretation of houses from images [1] and robotics [3, 2]). My talk is about a possibilities for relational integration of both relational action and vision. The real potential of relational representations is that states can *share* information with actions (e.g. parameters, or more specifically *objects*). Possibilities exist to define novel languages for *interactive experimentation* with relational abstraction levels in the context of both complex visual input and complex behavioral output. This includes new types of interactions – for example dealing with scarce human feedback, new types of experimentation – for example incorporating visual feedback and physical manipulation, and new types of abstraction levels – such as probabilistic programming languages. I will present first steps towards amore tight integration of relational vision and relational action for interactive [2] learning settings. In addition I present several new problem domains.

[2] See also the IJCAI Workshop on Machine Learning forInteractive Systems (http://mlis-workshop.org/2013/)

References

1 L. Antanas, M. van Otterlo, J. Oramas, T. Tuytelaars, and L. De Raedt. There are plenty of places like home: Using relational representations in hierarchy for distance-based image understanding. *Neurocomputing*, 2013; *in press* (http://dx.doi.org/10.1016/j.neucom.2012.10.037).

2 B. Moldovan, P. Moreno, and M. van Otterlo. On the use of probabilistic relational affordance models for sequential manipulation tasks in robotics. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2013.

3 B. Moldovan, P. Moreno, M. van Otterlo, J. Santos-Victor, and L. De Raedt. Learning relational affordance models for robots in multi-object manipulation tasks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4373–4378, 2012.

4 M. van Otterlo. Intensional dynamic programming: A Rosetta stone for structured dynamic programming. *Journal of Algorithms*, 64:169–191, 2009.

5 M. van Otterlo. *The Logic of Adaptive Behavior.* IOS Press, Amsterdam, The Netherlands, 2009.

6 M. van Otterlo. Solving relational and first-order logical Markov decision processes: A survey. In M. Wiering and M. van Otterlo, editors, *Reinforcement Learning: State-of-the-Art*, chapter 8. Springer, 2012.

7 M. Wiering and M. van Otterlo. *Reinforcement Learning: State-of-the-Art.* Springer, 2012.

3.31 Universal RL: applications and approximations

Joel Veness (University of Alberta – Edmonton, CA)

Joint work of Veness, Joel; Hutter, Marcus
Main reference J. Veness, "Universal RL: applications and approximations," 2013.
URL http://ewrl.files.wordpress.com/2012/10/ewrl11_submission_27.pdf

While the main ideas underlying Universal RL have existed for over a decadenow (see [1] for historical context), practical applications are only juststarting to emerge. In particular, the direct approximation introduced by Venesset al. [2, 3] was shown empirically to compare favorably to a number of other model-based RL techniques on small, partially observable environments with initially unknown, stochastic dynamics. Since then, a variety of additional techniques have been introduced that allow for the construction of far more sophisticated approximations. We present and review some of the mainideas that have the potential to lead to larger scale applications.

References

1 M. Hutter. *One decade of universal artificial intelligence.* In Theoretical Foundations of Artificial General Intelligence, pages 67–88. Atlantis Press, 2012.

2 J. Veness, K. S. Ng, M. Hutter, and D. Silver. Reinforcement Learning via AIXI Approximation. In AAAI, 2010.

3 J. Veness, K. S. Ng, M. Hutter, W. T. B. Uther, and D. Silver. A Monte-Carlo AIXI Approximation. Journal of Artificial Intelligence Research (JAIR), 40:95–142, 2011.

3.32 Learning and Reasoning with POMDPs in Robots

Jeremy Wyatt (University of Birmingham – Birmingham, GB)

Sequential decision making under state uncertainty is a well understood if intractable problem.In this talk I show various ways that approximate methods for belief state planning and learningcan be developed to solve practical robot problems including those that scale to high dimensional continuous state and action spaces, problems with incomplete information, and problems requiring real-time decision making.

4 Schedule

- Monday
  Morning
  - Self-introduction of participants
  - **Peter Sunehag:** Tutorial on exploration/exploitation

  Afternoon
  - **Tom Dietterich:** Solving Simulator-Defined MDPs for Natural Resource Management
  - **Csaba Szepesvari, Gergely Neu:** Online learning in Markov decision processes
  - **Ronald Ortner:** Continuous RL and restless bandits
  - **Martijn van Otterlo:** Relations Between Reinforcement Learning, Visual Input, Perception and Action
- Tuesday
  Morning
  - **M. Ghavamzadeh and A. Lazaric:** Tutorial on Statistical Learning Theory in RL and Approximate Dynamic Programming
  - **Gerard Neumann:** Hierarchical Learning of Motor Skills with Information-Theoretic Policy Search
  - **Doina Precup:** Methods for Bellman Error Basis Function construction
  - **Joëlle Pineau:** Reinforcement Learning using Kernel-Based Stochastic Factorization

  Afternoon
  - **Petar Kormushev:** Reinforcement Learning with Heterogeneous Policy Representations
  - **Mark B. Ring:** Continual learning
  - **Rich Sutton:** The quest for the ultimate TD algorithm
  - **Jan H Metzen:** Learning Skill Templates for Parameterized Tasks
  - **Robert Busa-Fekete:** Preference-based Evolutionary Direct Policy Search
  - **Orhan Sönmez:** Sequentially Interacting Markov Chain Monte Carlo Based Policy Iteration
- Wednesday
  Morning
  - **Marcus Hutter:** Tutorial on Universal RL
  - **Joel Veness:** Universal RL – Applications and Approximations
  - **Laurent Orseau:** Optimal Universal Explorative Agents
  - **Tor Lattimore:** Bayesian Reinforcement Learning + Exploration
  - **Laurent Orseau:** More realistic assumptions for RL

Afternoon
- **Scott Sanner:** Tutorial on Symbolic Dynamic Programming
- **Rico Jonschkowski:** Representation Learning for Reinforcement Learning in Robotics
- **Timothy Mann:** Theoretical Analysis of Planning with Options
- **Mohammad Ghavamzadeh:** SPSA based Actor-Critic Algorithm for Risk Sensitive Control
- **David Silver:** Deterministic Policy Gradients

- Thursday

  Morning
  - **Joëlle Pineau:** Tutorial on POMDPs
  - **Jeremy Wyatt:** Learning and Reasoning with POMDPs in Robots
  - **Scott Sanner:** Recent Advances in Symbolic Dynamic Programming for Hybrid MDPs and POMDPs
  - **Lutz Frommberger:** Some thoughts on Transfer Learning in RL – On States and Representation
  - **Will Uther:** Tree-based MDP – PSR
  - **Nils Siebel:** Neuro-Evolution

  Afternoon Hiking
- Friday

  Morning
  - **Christos Dimitrakakis:** ABC and cover-tree reinforcement learning
  - **Manuela Ruiz-Montiel:** Multi-objective Reinforcement Learning
  - **Ann Nowe:** Multi-Objective Reinforcement Learning
  - **Rico Jonschkowski:** Temporal Abstraction in Reinforcement Learning with Proximity Statistics
  - **Christos Dimitrakakis:** RL competition

Participants

- Peter Auer
Montan-Universität Leoben, AT
- Manuel Blum
Albert-Ludwigs-Universität Freiburg, DE
- Robert Busa-Fekete
Universität Marburg, DE
- Yann Chevaleyre
University of Paris North, FR
- Marc Deisenroth
TU Darmstadt, DE
- Thomas G. Dietterich
Oregon State University, US
- Christos Dimitrakakis
EPFL – Lausanne, CH
- Lutz Frommberger
Universität Bremen, DE
- Jens Garstka
FernUniversität in Hagen, DE
- Mohammad Ghavamzadeh
INRIA Nord Europe – Lille, FR
- Marcus Hutter
Australian National Univ., AU
- Rico Jonschkowski
TU Berlin, DE
- Petar Kormushev
Italian Institute of Technology – Genova, IT
- Tor Lattimore
Australian National Univ., AU
- Alessandro Lazaric
INRIA Nord Europe – Lille, FR
- Timothy Mann
Technion – Haifa, IL
- Jan Hendrik Metzen
Universität Bremen, DE
- Gergely Neu
Budapest University of Technology & Economics, HU
- Gerhard Neumann
TU Darmstadt, DE
- Ann Nowé
Free University of Brussels, BE
- Laurent Orseau
AgroParisTech – Paris, FR
- Ronald Ortner
Montan-Universität Leoben, AT
- Joëlle Pineau
McGill Univ. – Montreal, CA
- Doina Precup
McGill Univ. – Montreal, CA
- Mark B. Ring
Anaheim Hills, US
- Manuela Ruiz-Montiel
University of Malaga, ES
- Scott Sanner
NICTA – Canberra, AU
- Nils T. Siebel
Hochschule für Technik und Wirtschaft – Berlin, DE
- David Silver
University College London, GB
- Orhan Sönmez
Bogaziçi Univ. – Istanbul, TR
- Peter Sunehag
Australian National Univ., AU
- Richard S. Sutton
University of Alberta, CA
- Csaba Szepesvári
University of Alberta, CA
- William Uther
Google – Sydney, AU
- Martijn van Otterlo
Radboud Univ. Nijmegen, NL
- Joel Veness
University of Alberta, CA
- Jeremy L. Wyatt
University of Birmingham, GB

Report from Dagstuhl Seminar 13322

The Critical Internet Infrastructure

Edited by
Georg Carle[1], Jochen Schiller[2], Steve Uhlig[3], Walter Willinger[4], and Matthias Wählisch[2]

1 **TU München, DE,** carle@in.tum.de
2 **FU Berlin, DE,** {jochen.schiller,m.waehlisch}@fu-berlin.de
3 **Queen Mary University of London, GB,** steve@eecs.qmul.ac.uk
4 **AT&T Research – Florham Park, US**

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 13322 "The Critical Internet Infrastructure". The scope of the seminar includes three main topics, rethinking perspectives on the Internet backbone, methodologies to analyze the Internet structure, and paradigms overlaying IP connectivity. The results are based on fruitful discussions between people from the research and operational community.

Seminar 04.–09. August, 2013 – www.dagstuhl.de/13322
1998 ACM Subject Classification C.2.1 Network Architecture and Design, C.2.5 Local and Wide Area Network—Internet, C.2.6 Internetworking
Keywords and phrases Internet, Backbone, Internet Services, Critical Infrastructure
Digital Object Identifier 10.4230/DagRep.3.8.27
Edited in cooperation with Thomas C. Schmidt

1 Executive Summary

Georg Carle
Jochen Schiller
Steve Uhlig
Walter Willinger
Thomas C. Schmidt
Matthias Wählisch

The Internet was designed to offer open data transfer services on a planetary scale. However, its success has turned it into a mission-critical infrastructure of vital importance for most countries, businesses, and industries. The aim of this seminar is to bring together the research and network operator communities to discuss and analyze the Internet as a critical infrastructure. We will address the vulnerability of the Internet from a number of different angles (e.g., physical infrastructure, control plane, data plane, services, etc.), with an emphasis on the core transport infrastructure as well as the content delivery side. The seminar will contribute to a better understanding of the Internet as a system of interdependent elements and pursue extensions of current research perspectives to consider novel (and maybe unusual) approaches to studying the local or region-specific substrates as parts of the Internet's global ecosystem.

Rethinking Perspectives on the Internet Backbone

Analyzing the mutual impact between ASes, the vulnerability and efficiency of the backbone requires the identification of ASes and their role in mutual transit. In particular, stakeholders do not want to ground their internal data exchange on weak third parties. In Internet terms, AS interconnections between key players of a country should be part of a transparently visible Internet ecosystem. However, the Internet is a globally distributed network without boundaries, which makes the identification of locally relevant subparts hard. This seminar aims at being a platform to leverage new and uncommon research perspectives that go beyond the Internet backbone as a globally distributed system.

Methodologies to Analyze the Internet Structure

To analyze the Internet as a critical infrastructure, a clear picture is required about the kind and granularity of data needed to obtain relevant results and draw valid conclusions, even if the available dataset is restricted. Sampling and inference are common methods to assess the impact of the limited view on the real Internet. Current approaches to model the Internet backbone need to be revisited to reflect the Internet as critical infrastructure. The mapping of logical Internet nodes (ASes) to concrete entities (companies, points of presence etc.) as well as its annotation with meta data (e.g., administrative contact points) have been identified as important to cover the Internet structure from a non-technical perspective.

Paradigms Overlaying IP Connectivity

Delivering content to the end users is one of the main objectives of the Internet. In the early Internet, end users accessed content directly from a primary source. With the advent of CDNs this has changed. A single CDN operates as replication and distribution network for many content publishers, which brings data closer and more efficiently to end users. In fact, a very large portion of the current Internet content is maintained by only a limited number of CDNs, creating a limited competition in this area. Until now, this oligarchy has not been thoroughly studied, especially in the context of the Internet as a critical infrastructure.

Original Goals of the Seminar

The research questions to be pursued and answered include:

- How can we define and extract a locally relevant view of the globally distributed Internet?
- Which metrics are appropriate to measure the importance of Internet stakeholders and their mutual relationships?
- Which countermeasures and improvements are feasible to protect the Internet as critical infrastructure without narrowing its flexibility and openness?
- To what extent can we analyze the Internet structure in short time frames?
- What is the role of specific ASes for reliably interconnecting the Internet infrastructure of a country?
- How can we reveal weak transits and unintentionally strong dependencies between ASes and specific regions of the world?
- How can we predict Internet scale consequences of large scale problems (what-if-questions)?

The complexity of the Internet makes it equally complex to give complete answers to these questions. This seminar helped us to start *touching* the questions. During our discussions it was clear that it is not only important to continue the work on these challenges but that it is also worth to follow up with a more specific focus on measurement aspects.

Acknowledgments

The editors of this report would like to thank all participants for very fruitful and open-minded discussions! In particular, we thank the operators for sharing practical insights.

We gratefully acknowledge the Dagstuhl staff for helping on all administrative coordination, for their patience, and most importantly for providing an extremely inspiring environment.

2 Table of Contents

3 Overview of Talks

3.1 The impact of amended copyright acts to broadband traffic in Japan

Kenjiro Cho (Internet Initiative Japan Inc. – Tokyo, Japan)

This talk describes the impact of two copyright acts in Japan, which are effective since January 2010 and October 2012, respectively. The first act implements a download ban, making illegal content download illegal. This had an impact on the long-term traffic trend. The second act criminalized illegal downloads. The impact was only temporary. Our analysis is based on a clean and simple measurement setup.

3.2 Can you ping me now? – or – Measuring Mobile Networks

David Choffnes (Northeastern University – Boston, US)

Joint work of Choffnes, David; Mao, Morley; Zarifis, Kyriakos; Flach, Tobias; Nori, Srikanth; Katz-Bassett, Ethan; Govindan, Ramesh; Welsh, Matt; Hamon, Dominic; Feamster, Nick

Mobile networks are currently the fastest growing, most popular and least understood systems in today?s Internet ecosystem. Despite a need for performance improvement and policy transparency in this space, researchers currently struggle to measure, analyze, and optimize mobile networks.

To address this problem, we propose building Mobilyzer, an open platform for network measurement from mobile devices, to capture a detailed view of this complex and dynamic setting. By measuring mobile network performance directly and intelligently from devices, we will fill a critical gap: visibility from users? perspectives. First, I demonstrate the usefulness of such information by using a large dataset of network measurements from mobile devices to diagnose path inflation experienced by mobile users. Then I describe a mobile-network measurement platform that captures a continuous, broad view of mobile system interactions, annotated with contextual information (e.g., GPS location, signal strength and radio state) necessary to interpret the raw measurements. To facilitate research and foster innovation in this environment, our platform will be open to all researchers, support a flexible set of measurement techniques, it will transparently manage data collection and reporting, and the measurements will be publicly available and easily accessible.

3.3 Toward Realtime Visualization of Garbage

Alberto Dainotti (San Diego Supercomputer Center, US)

Joint work of Alistair, King; Dainotti, Alberto;

CAIDA, the Cooperative Association for Internet Data Analysis, investigates practical and theoretical aspects of the Internet. Using the CAIDA telescope, we monitor backscatter traffic to complement BGP measurements and to analyze country-level outages as well as routing incidents. Getting real-time insights into the gathered data is nearly impossible due to the large amount of volume. In this talk, we report about ongoing work on real-time visualization of very large data sets. We describe our pipeline-based architecture and present a live demo. We conclude with lessons learned and next steps.

3.4 Security and Attacks

Roland Dobbins (Arbor Networks – Singapore, Singapore)

This talk discusses two topics. The first part deals with network core infrastructure protection and best practices. It includes an overview about infrastructure protection aspects, explains the different forwarding planes in routers, and presents for router and network hardening. The second part of this talk identifies wireless Internet traffic trends and challenges including comparisons between wireline and wireless mobile broadband traffic.

3.5 Internet census taken by an illegal botnet. A qualitative analysis of the measurement data

Anja Feldmann (TU Berlin, DE)

This talk gives a qualitative analysis of the Internet census 2012, an anonymous port scanning of the Internet address space using insecure embedded devices. After a brief discussion of who was interested in the data, the talk contrasts claims made by the anonymous writer with the actual data set. The results were not not only unethically collected, but are also based on methodological flaws (e.g., uneven probing rates). The talk concludes that using the data for further studies is nearly impossible as insufficient meta and measurement data is provided.

3.6 Resilience of the Interdomain Routing System

Thomas Haeberlen (ENISA – Athens, GR)

We briefly give some points on what could be done to further improve the resilience of the interdomain routing, and present an approach with a slight change in perspective, which we are currently investigating in a project under ENISA's 2013 work programme

3.7 How I Will Measure Routes in 2014

Ethan Katz-Bassett (USC – Los Angeles, US)

Joint work of Katz-Bassett, Ethan; Calder, Matt; Zarifis, Kyriakos; Feamster, Nick; Cunha, Italo; Choffnes, Dave; Madhyastha, Harsha

In this talk we give a brief overview about past, present, and future work on how we measure routes. We explain the concept of the Transit Portal (or BGPMux) project, which lets researchers experiment with BGP in the wild by emulating an autonomous system. Transit Portal has been used in several analysis, e.g., for root cause analysis of BGP path changes. The second part of this talk deals with a unified platform for path queries. We discuss the need for single interface to access all sets of vantage points such as academic testbeds, enduser measurements, or route collectors.

References

1 E. Katz-Bassett, C. Scott, D. Choffnes, I. Cunha, V. Valancius, N. Feamster, H. Madhyastha, T. Anderson, A. Krishnamurthy. *LIFEGUARD: Practical Repair of Persistent Route Failures*. Proc. of ACM SIGCOMM, 2012
2 U. Javed, I. Cunha, D. Choffnes, E. Katz-Bassett, A. Krishnamurthy, T. Anderson. *PoiRoot: Investigating the Root Cause of Interdomain Path Changes* . Proc. of ACM SIGCOMM, 2013.

3.8 About ENISA. Security and resilience of the European communications networks

Rossella Mattioli (ENISA – Athens, GR)

ENISA, the European Union Agency for Network and Information Security, works on the improvement of network and information security in the European Union. ENISA gives advice on information security issues to national authorities, EU institutions, citizens, and business. It acts as a forum for sharing good NIS practices and facilitates information exchange and collaboration. In this talk, I will present background about ENISA and the NIS landscape. I will discuss the difference between Critical Infrastructure (CI) and Critical Information Infrastructure (CII). I will also consider EU legislation.

3.9 Configuration Complexity

Matthew Roughan (University of Adelaide, AU)

Network complexity is an important topic for network designers. The choice of design often comes down to "simple is better", but we lack metrics that quantify complexity. This talk is about defining complexity, at least for a limited domain – network configuration complexity. We argue that when the problem is decomposed correctly into network and protocol components we can use an analogue of Kolmogorov complexity to quantify how hard it is to define a network configuration, but there are still many unknowns. For instance, how to relate such a metric to the cost of operational tasks.

3.10 Internet PoP Level Maps and Beyond

Yuval Shavitt (Tel Aviv University, IL)

Joint work of Shavitt, Yuval; Zilberman, Noa

A large amount of monitoring and analysis research is devoted to the study of the Internet topology. There are several levels on which the Internet maps are presented, each level of abstraction is suitable for studying different aspects of the network. The most detailed level is the IP or router level, which represents separately each and every entity connected to the network. This level is far too detailed to suit practical purposes, and the large number of entities makes it very hard to handle. The coarsest level is the Autonomous System (AS) level. It is most commonly used to draw Internet maps, as it is relatively small (tens of thousands of ASes) and therefore relatively easy to handle: There is only one node for every AS, and may have only one edge between every pair of ASes. One limitation of using the AS level is that it cannot serve as tool to track the internet evolution, since AS sizes may differ by orders of magnitude. While a large AS can span an entire continent, a small one can serve a small community, yet both seem identical of the AS level map.

An interim level between the AS and the router graphs is the PoP level. Service providers tend to place multiple routers in a single location called a Point of Presence (PoP), which serves a certain area. A PoP is defined as a group of routers which belong to a single AS and are physically located at the same building or campus. Thus, for studying the Internet evolution and for many other tasks, PoP maps give a better level of aggregation than router level maps with minimal loss of information. PoP level graphs provide the ability to examine the size of each AS network by the number of physical co-locations and their connectivity instead of by the number of its routers and IP links. Points of presence can be annotated with geographical location, as well as information about the size of the PoP. Thus, using PoP level graphs it is possible to detect important nodes of the network and understand network dynamics as well as many more applications.

In the talk I present an algorithm for classification of IP addresses into PoPs and then present PoP level maps that are built by connecting the PoP nodes with edges that are aggregation of IP level links. I will then present algorithms for embedding this map in geography, namely assigning a geographic location to points of presence. I will present analysis of the different algorithm performance, and validation studies against ground truth data.

References

1 Lior Neudorfer, Yuval Shavitt, Noa Zilberman. *Improving AS relationship inference using PoPs*. Proc. of IEEE INFOCOM, IEEE Press, USA, 2013

2 Dima Feldman, Yuval Shavitt, Noa Zilberman. *A structural approach for PoP geo-location*. Computer Networks 56(3): pp. 1029-1040, 2012

3 Yuval Shavitt, Noa Zilberman. *Geographical Internet PoP Level Maps*. Proc. of TMA, Springer-Verlag, pp. 121-124, 2012

3.11 Economics

Bill Woodcock (PCH – San Francisco, US)

This talk discusses economic aspects of the critical Internet infrastructure from a cost perspective. It identifies critical infrastructure costs and how resiliency and defense affect these costs. Furthermore, this talk reflects 'perverse' incentives, zero tolerance vs. risk management, and deterrence versus mitigation versus retribution.

4 Working Groups

4.1 Control Plane Attack

The group identified the following vulnerable parts of the control plane: routing (BGP and IGP), DNS, MPLS, mobile packet control, higher layer control loops (e.g., CDN, VPN), and the so-called metaplane. The metaplane comprises authorized information about a third party, e.g., PKI in general and RPKI in particular for routing.

These parts can be attacked by misconfiguration and hacking. A major challenge are control loop problems as well as cascading failures. The operation of the current Internet is based on a highly complex system with several interacting components. In contrast to previous time where lower layer control loops were visible (e.g., BGP wedgies that lock into undesirable states), today there are much more higher layer control loops. A typical example are Content Delivery Networks (CDNs). ISPs itself do not need to implement traffic engineering because CDNs can very quickly shift traffic to other CDNs or ISPs. Under bad circumstances this may lead to slow oscillations. The group explained a scenario, in which an ISP changes its IGP. This results in large swing in BGP egress decision and thus causes traffic change, which finally will be countered by the CDN. Consequently, there is a significant relocation of traffic for the ISP. The ISP may try to rebalance load by using IGP. The whole process repeats as IGP operations and CDN traffic engineering interact with each other.

The group also illustrated cascading failures. One example is the Baofeng outage in 2009. Boafeng is a popular media player in China. The attack did not target the Boafeng service but the DNS infrastructure of Boafeng's registrar. In combination with a design flaw within the Boafeng client, which continuously started DNS requests, this led to a DNS amplification attack: In the first stage the DNS servers were not available due to the attacker.

In the second stage Boafeng clients started accidentally a DNS flood as they were not able to contact the originally originally DNS servers. These DNS queries flooded the network of China Telecom affecting several hundred millions of users for hours.

Control plane attacks are complicated. Even though an attack does not directly target the control plane, the attack may affect this layer due to many interacting protocols on different layers.

4.2 Government-Level Adversaries

The Internet is a critical infrastructure that should be protected by the government. However, it gives also new potential for misusing by the government to conduct attacks. The group analyzed the differential of power between attacker and defender. A government that attacks an individual will most likely be successful as end users lack sufficient resources. This perspective changes in case of a governmental adversary against a corporation or another government. Larger companies provision much more (Internet-based) resources to some state resources.

In the second part the group discussed how do governmental attacks characteristically differ from non-governmental attacks. The identified three items: (a) Government are more likely to make long-term investments and preparations; (b) governments are more likely to use cyber attacks as a component in a broader IO, which is itself more likely to be a component of a larger coordinated operation; (b) governments are more likely to apply classic espionage tradecraft in combination with cyber techniques.

A prominent example for a governmental adversary is TOR, The Onion Router. TOR was designed as 'circumvention' tool that allows for hiding information. A widespread deployment of TOR scatters data in such a way, which makes decryption costly. This is not only useful for the common public but also helps to cover agents. Unfortunately, Moore's Law gradually democratised TOR. It allows more intelligence services to compete for control, causing the US government to lose interest in TOR and support the development of new competing platforms.

Finally, the transition from certificates to DNS-based Authentication of Named Entities (DANE) has been discussed briefly. The deployment of X.509 PKI gives certificate issuers significant power. This makes certificate authorities very attractive for governmental attacks. DANE (cf., RFC 6698) vastly reduces the number of potential compromisers, but arguably gives governments more direct ability to attack their own citizens and corporations, since governments often control the signing of their country-code domain (ccTLD).

4.3 Disaster Recovery

The group started the discussion based on two talks. The first talk was given by Kenjiro Cho who presented insights from Japan Earthquake with respect to the impact on traffic and routing observed by IIJ, a local ISP. The second talk was given by Randy Bush who presented the DUMBO project, which is about using MANET for disaster management, exemplified on recovering from South-East Asian Tsunamis. The group agreed that the Internet matters after a disaster; e.g., for coordinating emergency responses or the dissemination of food, water, etc. Furthermore, the role of the Internet after a disaster is to provide information. The communication service is not necessarily email, because it needs to reach the broader public to prevent panic.

The main outcome of the discussion was that the *culture* is surprisingly important in case of disaster recovery. In the example of the Japan Earthquake, IIJ had enough spare capacities to deal with rerouting all network traffic. This high over-provisioning was by intention and is different in other countries.

The DUMBO project illustrated nicely another aspect, the consideration of the local environment. The project partners used elephants to carry equipment in South-East Asia because it is both, culturally appropriate and handy after a disaster.

Bill Woodcock gave an example where not only the cultural of a society but also of organizations need to be considered. After the Haiti earthquake in 2010, ISPs had 90 % of the network up using wireless technology. However, However, NGOs that arrived soon after to help accidentally broke network communication by taking over the wireless spectrum.

4.4 Mapping (Inter-)National Infrastructure

The group discussed the localization of the Internet infrastructure with respect to geographic, political, sociological, or organisational aspects. Perspectives are national (i.e., seen from within countries) or international (i.e., between countries). However, rather unclear is what entities can/shall be mapped, and what can(not) be done on a technical level. The group identified four building blocks of interest:

1. Provisioning of infrastructure.
2. Civil organization and business.
3. Relations to legislature.
4. Aspects of Internet use.

As major problem in order to map the infrastructure, missing public information has been identified. The level of public information depends on the country and regulatory requirements. In Poland, for example, complete maps of cables are publicly available. In contrast to this, in Germany operators are required to provide certain information about location etc. However, this information is treated confidentially.

Another challenge in evaluating the quality of public data is the question of intended purpose while publishing the information, in particular in the business context. Searching the web for public data usually brings more data to light than expected (e.g., on operator websites). But what is the motivation for a company to publish such data? The data maybe highly specific to distract from the actual (desired) insights.

Currently there are no central databases that can provide a complete and correct view on all aspects related to the Internet infrastructure. In fact, we expect that such a view is unlikely to be available soon. There are on-going activities to build new and to refine existing mapping approaches. Evaluating the quality of the solutions depends significantly on the ground truth in the data.

4.5 Non-Adversarial Threats to Availability

During this break-out session the group discussed threats that have not been introduced by an attacker. Several concrete examples have been identified. For example, (and maybe the most important) the lack of power. Many critical data centers etc. are protected by uninterruptible power supplies. A proper working of those systems requires regular maintenance (e.g., test of generators), which is surprisingly often missed.

Another example is the lack of hardcopy critical documentation and contact information, or failure to train employees to find such, or failure to keep such documentation up-to-date. Accessing this information should not be bound to the correct function of the technical system (e.g., authorization).

A common problem is the misunderstanding of redundancy. The group illustrated this by the following anecdote: Microsoft was buying 150 % of all of the capacity into Seattle, mistakenly thinking they were buying three redundant DS3s of Internet connectivity, at a time when there were only two data DS3s leaving the city. It was also observed that several people believe that they are creating redundancy, while they are actually putting components in serial (increasing fragility) rather than in parallel. Or they put components in parallel, but in such a way that the operation of one is in some subtle way dependent upon the operation of the other, which amounts to the same thing.

The main outcome of this session was that non-adversarial threats are multi-dimensional including technical and non-technical aspects.

4.6 A Unified Interface for Measurements

This group discussed the potential of a unified access to the plethora of existing measurement data sets. The group provided insights from relevant past EU projects and found that involvement of people with long-term stakes in essential, simple approaches are better than complex ones, and differentiation between “small” and “large” data queries is critical.

One important design decision in building the related data infrastructure is the question of a central versus distributed repository. Central repositories need periodic data updates but may support more subtle queries. On the other hand, a distributed repository lowers the barrier for entry and is easier to maintain. On idea was to implement a gradual deployment, i.e., start with a distributed system and migrate to a central system later. As hosting options M-Lab, ISI, and RIPE have been identified.

Participants

- Bernhard Ager
ETH Zürich, CH
- Lothar Braun
TU München, DE
- Randy Bush
Internet Initiative Japan Inc. – Tokyo, JP
- Georg Carle
TU München, DE
- Nikolaos Chatzis
TU Berlin, DE
- Kenjiro Cho
Internet Initiative Japan Inc. – Tokyo, JP
- David Choffnes
Northeastern University – Boston, US
- Alberto Dainotti
San Diego Supercomputer Center, US
- Roland Dobbins
Arbor Networks – Singapore, SG
- Anja Feldmann
TU Berlin, DE
- Timothy G. Griffin
University of Cambridge, GB
- Thomas Häberlen
ENISA – Athens, GR
- Ethan Katz-Bassett
USC – Los Angeles, US
- Stefan Katzenbeisser
TU Darmstadt, DE
- Rossella Mattioli
ENISA – Athens, GR
- Matthew Roughan
University of Adelaide, AU
- Jochen Schiller
FU Berlin, DE
- Johann Schlamp
TU München, DE
- Thomas C. Schmidt
HAW – Hamburg, DE
- Yuval Shavitt
Tel Aviv University, IL
- Georgios Smaragdakis
T-Labs/TU Berlin
- Rade Stanojevic
Telefónica Res. – Barcelona, ES
- Steve Uhlig
Queen Mary University of London, GB
- Matthias Wählisch
FU Berlin, DE
- Walter Willinger
AT&T Labs Research – Florham Park, US
- Bill Woodcock
PCH – San Francisco, US

Report from Dagstuhl Seminar 13331

Exponential Algorithms: Algorithms and Complexity Beyond Polynomial Time

Edited by

Thore Husfeldt[1], Ramamohan Paturi[2], Gregory B. Sorkin[3], and Ryan Williams[4]

1 **IT University of Copenhagen, DK,** thore.husfeldt@gmail.com
2 **UC San Diego, US,** paturi@cs.ucsd.edu
3 **London School of Economics, GB,** G.B.Sorkin@lse.ac.uk
4 **Stanford University, US,** rrwilliams@gmail.com

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 13331 "Exponential Algorithms: Algorithms and Complexity Beyond Polynomial Time". Problems are often solved in practice by algorithms with worst-case exponential time complexity. It is of interest to find the fastest algorithms for a given problem, be it polynomial, exponential, or something in between. The focus of the seminar is on finer-grained notions of complexity than NP-completeness and on understanding the exact complexities of problems. The report provides a rationale for the workshop and chronicles the presentations at the workshop. The report notes the progress on the open problems posed at the past workshops on the same topic. It also reports a collection of results that cite the presentations at the previous seminar. The docoument presents the collection of the abstracts of the results presented at the seminar. It also presents a compendium of open problems.

Seminar 11.–16. August, 2013 – www.dagstuhl.de/13331
1998 ACM Subject Classification F.1.3 Complexity Measures and Classes, F.2 Analysis of Algorithms and Problem Complexity
Keywords and phrases Algorithms, exponential time algorithms, exact algorithms, computational complexity, satisfiability
Digital Object Identifier 10.4230/DagRep.3.8.40
Edited in cooperation with Dominik Scheder

1 Executive Summary

Thore Husfeldt
Ramamohan Paturi
Gregory Sorkin
Ryan Williams

Background

Computational complexity has demonstrated that thousands of important computational problems, spanning the sciences, are intimately linked: either they all have polynomial time algorithms, or none does. Nearly all researchers believe that P $\neq$ NP, and that these problems do not all have low time complexity. However, they must be solved, one way or another, which means relaxing the requirements for “solving” a problem. One natural requirement to relax is the running time. Problems are often solved in practice by algorithms with worst-case exponential time complexity. It is of interest to find the *fastest* algorithm for a given problem, be it polynomial, exponential, or something in between.

This relaxation has revealed a finer-grained notion of problem complexity than NP-completeness. By definition all NP-complete problems are equivalent as far as the existence of polynomial time algorithms is concerned. However, the exact time complexities of these problems can be very different, just as their best approximation ratios can vary.

Algorithms for satisfiability represent well the progress in the field and the questions that arise. The theory of NP-completeness says that the Circuit Sat problem and 3-Sat are polynomial time equivalent. However, from the exact, exponential time perspective, the two problems look radically different.

For 3-Sat (and k-Sat in general), algorithms faster than the exhaustive search of 2^n assignments have been known for many years and are continually improved. The analysis of the randomized PPSZ algorithm for 3-Sat has recently been improved to $O(1.308^n)$, so currently the best known algorithm for this problem is also very simple. The best known deterministic algorithm runs in time $O(1.331^n)$, and is obtained by derandomizing earlier local search algorithms. A very natural DPLL-type algorithm for Formula Sat in fact has good performance on linear size formulas. All of these results represent major conceptual contributions.

No such progress has been made for general Circuit Sat. In fact, such results would have major implications in circuit complexity: even a $1.99^n poly(m)$ time algorithm for satisfiability of circuits with n inputs and m gates would imply exponential size *lower bounds* for solving problems with circuits. Between 3-Sat and Circuit Sat, there are also intermediate problems such as CNF-Sat that have resisted all efforts to produce an $O(1.99^n)$ time algorithm.

The basic algorithmic techniques to avoid exhaustive search are now consolidated in the field’s first textbook, (Fomin and Kratsch, *Exact Exponential Algorithms*, Springer 2010) though they are still being extended and refined. For example, there is now a general framework for making various exponential time dynamic programming algorithms, such as standard algorithms for Knapsack and Subset Sum, run in polynomial space. The fast zeta transform, which plays a central role in the implementation of inclusion-exclusion algorithms, continues to be actively researched. And “measure-and-conquer” methods for analyzing branching/backtracking algorithms continue to be enhanced.

However, many other powerful techniques have been explored only recently. One idea is to find combinatorial structures (such as matchings) by looking for corresponding algebraic objects (such as polynomials). The idea dates to Edmonds if not Tutte, but was introduced by Koutis for exponential time path and packing problems, leading to an $2^k poly(n)$ algorithm to find a k-path in a graph and a breakthrough $O(1.67^n)$ time algorithm for finding a Hamiltonian path, improving the 50-year-old previously best algorithm.

Other open problems in the field have been attacked by intricate, dedicated analyses; for example, there is now an algorithm for scheduling partially ordered jobs in $O((2-\epsilon)^n)$ time.

Parameterized complexity is a closely related field that also investigates exponential time computation. Fundamentally, the field is interested in the dichotomy between algorithms that admit running times of the form $f(k)poly(n)$ (called fixed-parameter tractability) and those that do not, leading to qualitative hardness notions like $W[1]$-hardness. This field continues to make great progress, with the parameterized tractability of many fundamental problems just being discovered. Just recently the first fixed-parameter algorithms for finding topological minors and the multi-cut problem have been found.

However, many recent results in that area are interested in determining (typically exponential) growth rate of the function f, instead of just establishing its existence. For example, a recent, very successful focus of parameterized complexity is the existence of problem *kernels* of polynomial size, or their absence under assumptions from classical computational complexity. In another direction, very strong lower bounds for algorithms parameterized by treewidth can now be shown under hypotheses about the exponential time complexity of Sat.

A quantitative theory of computational complexity of hard problems would address questions like why it is that 3-Sat can be solved in 1.4^n but CNF-Sat apparently cannot be solve. Ideally, we could hope for a characterization of the exact complexity of NP-complete problems, perhaps under some plausible assumptions. There is a growing body of work on the exact complexity of NP-complete problems which draws heavily from parameterized complexity theory. The *Exponential Time Hypothesis* (ETH), which posits that 3-Sat cannot be solved in $2^{o(n)}$ time, has given a strong explanatory framework for why some classes of problems admit improved algorithms while others are resistant. The results surrounding ETH show that if 3-Sat could be solved in subexponential time, then many other NP problems would also have subexponential algorithms.

Another compelling conjecture is the *Strong Exponential Time Hypothesis* (SETH) that CNF Satisfiability cannot be solved in 1.999^n time on formulas with n variables and cn clauses (for sufficiently large c). SETH has implications for k-Sat, other graph problems, and parameterized computation. There is less consensus about the truth of SETH; nevertheless, studying its implications will help better understand what makes CNF so difficult. A counting version of the hypothesis, #ETH, has recently been introduced to study the exponential time complexity of counting problems, such as the permanent and the Tutte polynomial.

Connections to other fields are being discovered as well, such as the importance of exponential time algorithms to the study of lower bounds in circuit complexity, as mentioned above.

For another example, a celebrated recent result in the complexity of approximation algorithms exhibits an $\exp(O(n^{\varepsilon}))$ time approximation algorithm for Khot's Unique Games problem. This suggests that approximating unique games is a significantly easier task than solving NP-hard problems such as 3-Sat. The key to the algorithm is a new type of graph decomposition based on spectral methods. This decomposition method may well lead to more developments in exponential algorithms.

Furthermore, there are surprising connections between SETH and various other well-studied questions from other areas such as communication complexity and the 3-Sum hypothesis used in computational geometry and data structures. The instance compressibility notion introduced in the study of kernelisation turns out to be connected to the construction of hash functions.

These results show that increased attention to exponential time algorithms leads to progress of the highest caliber in well-established areas of the theory of computation.

2 Table of Contents

3 About the Meeting

The meeting was attended by 42 researchers, slightly smaller number than the number anticipated due to a few last-minute cancellations. The organizers are grateful to all who came, and regret that – due to a high acceptance rate – others who would have contributed could not be invited. The participants came all around the globe, predominantly from Europe and with a good showing from Germany and US: AU 1, AT 1, CH 1, DE 6, DK 3, FI 2, FR 2, GB 3, HU 1, IN 1, JP 3, NL 2, NO 3, PL 2, RU 1, SE 1, US 9.

Attendees include eight graduate students or postdoctoral fellows and eleven faculty who are at the beginning of their academic career. The meeting has a fair representation of researchers from each of the three fields, exact exponential algorithms, parameterized complexity, and computational complexity.

Almost all the talks were 45 minutes long. The after-lunch period was left free for informal discussions and small working groups, with talks again between 4 PM and dinner time. There were three open problem sessions, each taking up 30 minutes. These sessions brought out 19 open problems which were described in a latter section. Wednesday included traditional outing with the options of a long hike and a long run.

R. Williams started the session with an overview of the Exponential Time Hypothesis (ETH) and the Strong Exponential Time Hypothesis (SETH) discussing the motivation for the hypotheses and their many consequences. Bodlaender showed how algebraic methods can be used to obtain faster algorithms for several problems on tree decompositions with running time that is single exponential in the treewidth and linear in the number of vertices. Kulikov presented an improved exponential time algorithm for a special case of the classical problem, shortest superstring. Kowalik talked about fast self-reducibility with the number of oracle calls linear in the size of the solution. Wahlström presented a surprising result that gives a polynomial compression for the Steiner cycle problem.

On Tuesday, several results were presented concerning improved algorithms for restricted circuit satisfiability problems.Tamaki talked about improved algorithm for formulas over full binary basis. Seto presented an improved algorithm for sparse instances of MaxSat. Schneider presented an algorithm for depth-2 threshold circuit satisfiability. Santhanam presented new algorithms for QBF satisfiability. Hertli presented a intriguing result that raises the hope for beating the current best algorithm for 3-SAT. In contrast, Nederlof presented a series of problems whose running time (in terms of the constant in the exponent) cannot be improved if SETH holds. The set of results presented on Tuesday enlarge our understanding of the possibility of improved algorithms for satisfiability as a function of the expressive power of the circuit class.

On Wednesday, Lokshtanov presented a nice account of the work on efficient computation of representative sets and its applications to exact algorithms and kernelisation. Lovett talked about the long-standing log rank conjecture in communication complexity and the recent progress he made. Drucker presented a new line of work which shows that the question of efficient algorithms for 3-SAT in probabilistic polynomial time can be tied to the relationship between NP and co-NP. Steurer talked about approximability in the region between polynomial and exponential time and how the results in this area an be unified using the sum-of-squares method.

On Thursday, several new algorithmic results were presented: Marx presented a subexponential time algorithm for the subset TSP problem for planar graphs. Saurabh presented an improved algorithm for the multicut problem. D. Kratsch presented an algorithm for the jumpnumber problem beating the 2^n upperbound. S. Kratsch showed how one can use matrix methods to obtain deterministic algorithms for problems such as Hamiltonian Cycle and Steiner Tree. Also Heggernes and Kaski presented results on enumeration of graph classes and counting thin subgraphs respectively. A highlight on Thursday was the result presented by V. Williams who showed that approximating the graph diameter beyond the 3/2 ratio is not possible unless SETH fails.

On Friday, Eberfeld talked about the development of a theory of parameterized space complexity. In addition, three algorithmic results were presented: Gaspers presented techniques to improve Measure and Conquer analyses and a faster algorithm for Max-2-CSP. Cygan presented a faster algorithm for the TSP problem when the average degree is bounded. Finally, Austrin presented improved time-space tradeoffs for the subset sum problem.

Several open problems from previous Dagstuhl Seminars (seminars 08431 and 10441) have been resolved. Ryan Williams asked about improved satisfiability algorithms for depth-2 threshold circuits in the seminar 10441. The question has been partially answered by Impagliazzo, Paturi and Schneider and presented in the seminar on Tuesday morning. Björklund, Kaski and Kowalik resolved an open problem posed by Koutis in the seminar 10441 regarding k-vector coloring. Binkele-Raible et al. resolved an open question posed by van Rooij in the seminar 08431 by providing an improved algorithm for computing irredundancy numbers. Heggernes, Kratsch, Lokshtanov, Raman and Saurabh showed that partitioning a permutation into monotone sequences is fixed parameter tractable when the parameter is the number of such sequences resolving an open problem stated by Kratsch in the seminar 08431.

We also found several research results which credit the seminar 10441. Junosza-Szaniawski, Kratocvhíl, Liedloff, Rossmanith, and Rzażewski cite the presentation by Rossmanith at seminar 10441 in their work on fast algorithms for $L(2,1)$-labeling of graphs. Björklund, Kaski, and Kowalik in their work on optimal graph motifs cite the presentation given by Koutis at the seminar 10441. Van Rooij generously credits the impact of the seminars 08431 and 10441 on his work in his PhD thesis. Dantsin and Hirsch's paper published in SAT 2011 notes that the work was done in part during the seminar 10441.

4 Organization

The seminar was organized by

Thore Husfeldt, IT University of Copenhagen, DK, thore@itu.dk
Ramamohan Paturi, University of California, San Diego, US, paturi@cs.ucsd.edu
Gregory Sorkin, London School of Economics, London, GB, G.B.Sorkin@lse.ac.uk
Ryan Williams, Stanford University, US, rrwilliams@gmail.com

Abstracts and open problems are being compiled and edited by

Dominik Scheder, Aarhus University, DK, and Simons Institute for the Theory of Computing, UC Berkeley, US, dominik.scheder@gmail.com,

who is also assisting in preparation of this final report.

We are grateful to the Dagstuhl personnel for their helpfulness and expertise, making the meeting smooth-running, pleasurable, productive, and easy to organize.

5 Overview of Talks

5.1 Space-Time Tradeoffs for Subset Sum

Per Austrin (KTH Stockholm, SE)

Joint work of Austrin, Per; Kaski, Petteri; Koivisto, Mikko; Määttä, Jussi
Main reference P. Austrin, P. Kaski, M. Koivisto, J. Määttä, "Space-Time Tradeoffs for Subset Sum: An Improved Worst Case Algorithm," arXiv:1303.0609v1 [cs.DS], 2013.
URL http://arxiv.org/abs/1303.0609v1

I discuss recent work on improved Space-Time Tradeoffs for the Subset Sum Problem.

5.2 Faster Algorithms on Tree Decompositions with Representative Sets: Expressibility and Experimentation

Hans L. Bodlaender (Utrecht University, NL)

Joint work of Bodlaender, Hans L.; Fafiani, Stefan; Kratsch, Stefan; Nederlof, Jesper

Recent work [1] (see also [2]) shows that algebraic methods can be used to obtain faster algorithms for several problems on tree decompositions, in particular for problems with connectivity requirements. The main underlying technique is to bring back a set of characteristics of partial solutions to a representative set. In this talk, we discuss a recent implementation of the method for the Steiner Tree problem [3], and a general characterization of a large set of problems that all allow such algorithms, running in time that is single exponential in the treewidth and linear in n.

References

1 Hans L. Bodlaender, Marek Cygan, Stefan Kratsch, Jesper Nederlof: Deterministic Single Exponential Time Algorithms for Connectivity Problems Parameterized by Treewidth. ICALP (1) 2013: 196-207.
2 Marek Cygan, Stefan Kratsch, Jesper Nederlof: Fast Hamiltonicity checking via bases of perfect matchings. STOC 2013: 301-310.
3 Fafianie, Stefan; Bodlaender, Hans L.; Nederlof, Jesper. Speeding-up Dynamic Programming with Representative Sets – An Experimental Evaluation of Algorithms for Steiner Tree on Tree Decompositions. arXiv:1305.7448, 2013. To appear in proceedings IPEC 2013.

5.3 TSP and counting perfect matchings in bounded average degree graphs

Marek Cygan (University of Warsaw, PL)

Joint work of Cygan, Marek; Pilipczuk, Marcin
Main reference M. Cygan, M. Pilipczuk, "Faster Exponential-Time Algorithms in Graphs of Bounded Average Degree," in Proc. of the 40th Int'l Colloquium on Automata, Languages, and Programming (ICALP'13), LNCS, Vol. 7965, pp. 364–375, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-39206-1_31

We first show that the Traveling Salesman Problem in an n-vertex graph with average degree bounded by d can be solved in $O^*(2^{(1-\epsilon_d)n})$ time and exponential space for a constant ϵ_d depending only on d. Thus, we generalize the recent results of Björklund et al. [TALG 2012] on graphs of bounded degree. Then, we move to the problem of counting perfect matchings in a graph. We first present a simple algorithm for counting perfect matchings in an n-vertex graph in $O^*(2^{n/2})$ time and polynomial space; our algorithm matches the complexity bounds of the algorithm of Björklund [SODA 2012], but relies on inclusion-exclusion principle instead of algebraic transformations. Building upon this result, we show that the number of perfect matchings in an n-vertex graph with average degree bounded by d can be computed in $O^*(2^{(1-\epsilon_{2d})n/2})$ time and exponential space, where ϵ_{2d} is the constant obtained by us for the Traveling Salesman Problem in graphs of average degree at most $2d$. Moreover we obtain a simple algorithm that computes a permanent of an $n \times n$ matrix over an arbitrary commutative ring with at most dn non-zero entries using $O^*(2^{(1-1/(3.55d))n})$ time and ring operations, improving and simplifying the recent result of Izumi and Wadayama [FOCS 2012].

5.4 Nondeterministic Direct Product Reductions and the Success Probability of SAT Solvers

Andrew Drucker (MIT, US)

Main reference A. Drucker, "Nondeterministic Direct Product Reductions and the Success Probability of SAT Solvers," to appear in Proc. of IEEE FOCS 2013.

In this talk I will describe nondeterministic reductions which yield new direct product theorems (DPTs) for Boolean circuits. In our theorems one assumes that a function F is "mildly hard" against *nondeterministic* circuits of some size $s(n)$, and concludes that the t-fold direct product F^t is "extremely hard" against probabilistic circuits of only polynomially smaller size $s'(n)$. The main advantage of these results compared with previous DPTs is the strength of the size bound in our conclusion. As an application, we show that if NP is not in coNP/poly then, for every PPT algorithm attempting to produce satisfying assignments to Boolean formulas, there are infinitely many satisfiable instances on which the algorithm's success probability is nearly-exponentially small. This furthers a project of Paturi and Pudlák [STOC'10].

5.5 Space in Parameterized Complexity

Michael Elberfeld (ICSI – Berkeley, US)

Joint work of Elberfeld, Michael; Stockhusen, Christoph; Tantau, Till
Main reference M. Elberfeld, C. Stockhusen, T. Tantau, "On the Space Complexity of Parameterized Problems," in Proc. of the 7th Int'l Symp. on Parameterized and Exact Computation (IPEC'12), LNCS, Vol. 7535, pp. 206–217, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-33293-7_20

Besides running time, the (memory) space available severely limits the range of computational problems that can be solved efficiently. Classical computational complexity provides several ways to classify the space complexity of problems using space bounds or additional simultaneous time bounds. The talk reports on an ongoing effort to develop a theory of parameterized space complexity and algorithms that is based on (1) standard parameterizations of classical space complexity classes, (2) space-efficient notions of bounded fixed-parameter tractability, and (3) parameter-based simultaneous resource bounds. Here I will focus on the first aspect and talk about parameterizations of deterministic logarithmic space; and how they can be used to give new insights into the complexity of well-studied parameterized problems. The talk details results from a joint work with Christoph Stockhusen and Till Tantau published in the proceedings of IPEC 2012 as well as surveys results from the literature to give an overview of the theory developed so far.

5.6 Separate, Measure, and Conquer

Serge Gaspers (UNSW – Sydney, AU)

Joint work of Gaspers, Serge; Sorkin, Gregory B.

Graph separators have been used extensively to design subexponential time algorithms for graph classes and to guide heuristics in practical solvers. We present a method to take advantage of linear-sized separators to improve Measure and Conquer analyses. As case analyses, we give faster polynomial-space algorithms for Max 2-CSP and # Dominating Set.

5.7 Enumeration in Graph Classes

Pinar Heggernes (University of Bergen, NO)

Enumerating, counting, and determining the maximum number of various objects in graphs have long been established as important areas within graph theory and graph algorithms. As the number of enumerated objects is very often exponential in the size of the input graph, enumeration algorithms fall into two categories depending on their running time: those whose running time is measured in the size of the input, and those whose running time is measured in the size of the output. Based on this, we concentrate on the following two types of algorithms. 1. Exact exponential time algorithms. The design of these algorithms is mainly based on recursive branching. The running time is a function of the size of the input graph, and very often it also gives an upper bound on the number of enumerated objects any graph can have. 2. Output polynomial algorithms. The running time of these algorithms is polynomial in the number of the enumerated objects that the input graph actually contains. Some of these algorithms have even better running times in form of incremental polynomial or polynomial delay, depending on the time the algorithm spends between each consecutive object that is output. The methods for designing the two types of algorithms are usually quite different. Common to both approaches is that efforts have traditionally mainly been concentrated on arbitray graphs, whereas graphs with particular structure have largely been left unattended. In this talk we look at enumeration of objects in graphs with special structure. In particular, we focus on enumerating minimal dominating sets in various graph classes.

Algorithms of type 1: The number of minimal dominating sets that any graph on n vertices can have is known to be at most 1.7159^n. This upper bound might not be tight, since no examples of graphs with 1.5705^n or more minimal dominating sets are known. For several classes of graphs, like chordal, split, and proper interval graphs, we substantially improve the upper bound on the number of minimal dominating sets. At the same time, we give algorithms for enumerating all minimal dominating sets, where the running time of each algorithm is within a polynomial factor of the proved upper bound for the graph class in question. In some cases, we provide examples of graphs containing the maximum possible number of minimal dominating sets for graphs in that class, thereby showing the corresponding upper bounds to be tight.

Algorithms of type 2: Enumeration of minimal dominating sets in graphs has very recently been shown to be equivalent to enumeration of minimal transversals in hypergraphs. The question whether the minimal transversals of a hypergraph can be enumerated in output polynomial time is a fundamental and challenging question; it has been open for several decades and has triggered extensive research. We show that all minimal dominating sets of a line graph can be generated in incremental polynomial, and consequently output polynomial, time. We are able to improve the delay further on line graphs of bipartite graphs. Finally we show that our method is also efficient on graphs of large girth, resulting in an incremental polynomial time algorithm to enumerate the minimal dominating sets of graphs of girth at least 7. The presentation is based on joint works with Jean-François Couturier, Pim van 't Hof, and Dieter Kratsch [1], and with Petr Golovach, Dieter Kratsch, and Yngve Villanger [2].

References

1. J-F. Couturier, P. Heggernes, P. van 't Hof, and D. Kratsch. Minimal dominating sets in graph classes: Combinatorial bounds and enumeration. Theoretical Computer Science 487: 82–94 (2013).
2. P. Golovach, P. Heggernes, D. Kratsch, and Y. Villanger. An incremental polynomial time algorithm to enumerate all minimal edge dominating sets. ICALP 2013, Lecture Notes in Computer Science 7965: 485–496 (2013).

5.8 A Faster Algorithm For Unique 3-SAT

Timon Hertli (ETH Zürich, CH)

The PPSZ algorithm by Paturi, Pudlák, Saks, and Zane (FOCS 1998) is the fastest known algorithm for Unique k-SAT. We give an improved algorithm with exponentially faster bounds for Unique 3-SAT. We show that worst-case formulas for PPSZ are either sparse or allow us to extract nontrivial information. In the former case, we use an algorithm by Wahlström, in the latter case, we improve the very beginning of PPSZ.

5.9 Counting thin subgraphs via packings faster than meet-in-the-middle time

Petteri Kaski (Aalto University, FI)

Joint work of Björklund, Andreas; Kaski, Petteri; Kowalik, Łukasz
Main reference A. Björklund, P. Kaski, Ł. Kowalik, "Counting thin subgraphs via packings faster than meet-in-the-middle time," arXiv:1306.4111v1 [cs.DS], 2013; to appear in Proc. of the ACM-SIAM Symp. on Discrete Algorithms (SODA'14, 5–7 January 2014, Portland, OR).
URL http://arxiv.org/abs/1306.4111

Vassilevska and Williams (STOC 2009) showed how to count simple paths on k vertices and matchings on $k/2$ edges in an n-vertex graph in time $n^{k/2+O(1)}$. In the same year, two different algorithms with the same runtime were given by Koutis and Williams (ICALP 2009), and Björklund *et al.* (ESA 2009), via $n^{st/2+O(1)}$-time algorithms for counting t-tuples of pairwise disjoint sets drawn from a given family of s-sized subsets of an n-element universe. Shortly afterwards, Alon and Gutner (TALG 2010) showed that these problems have $\Omega(n^{\lfloor st/2 \rfloor})$ and $\Omega(n^{\lfloor k/2 \rfloor})$ lower bounds when counting by color coding. Here we show that one can do better, namely, we show that the "meet-in-the- middle" exponent $st/2$ can be beaten and give an algorithm that counts in time $n^{0.4547st+O(1)}$ for t a multiple of three. This implies algorithms for counting occurrences of a fixed subgraph on k vertices and pathwidth $p \ll k$ in an n-vertex graph in $n^{0.4547k+2p+O(1)}$ time, improving on the three mentioned algorithms for paths and matchings, and circumventing the color-coding lower bound.

5.10 Fast Extraction and Listing of Witnesses Using a Decision Oracle

Łukasz Kowalik (University of Warsaw, PL)

Joint work of Björklund, Andreas; Kaski, Petteri; Kowalik, Łukasz

The gist of many (NP-)hard combinatorial problems is to decide whether a universe of n elements contains a witness consisting of k elements that match some prescribed pattern. Example: find a k-path in an n-vertex graph. The state-of-art results provide a probabilistic, one-sided error oracles for testing whether such a k-path exists and then the pattern itself can be found by self-reducibility. We show that the self-reducibility can be done very efficiently, namely with $O(k)$ running time overhead (or in $O(k \log n)$ queries in the more abstract oracle setting). We also investigate the task of listing all witnesses: this can be done with $O(k(\log n + \log s))$ delay between two successive witnesses, where s is the total number of witnesses (unknown to the algorithm).

5.11 The jumpnumber problem: exact and parameterized

Dieter Kratsch (University of Metz, FR)

The Jump Number problem asks to find a linear extension of a given partially ordered set that minimizes the total number of jumps, i.e., the total number of consecutive pairs of elements that are incomparable originally. The problem is known to be NP-complete even on posets of height one and on interval orders. It has also been shown to be fixed-parameter tractable. Finally, the Jump Number problem can be solved in time $O^*(2^n)$ by dynamic programming. In this talk we present an exact algorithm to solve Jump Number in $O(1.8638^n)$ time. We also show that the Jump Number problem on interval orders can be solved by an $O(1.7593^n)$ time algorithm, and prove fixed-parameter tractability in terms of width w by an $O^*(2^w)$ time algorithm. Furthermore, we give an almost-linear kernel for Jump Number on interval orders when parameterized by solution size.

5.12 Applications of matrix rank for faster dynamic programming on tree and path decompositions

Stefan Kratsch (TU Berlin, DE)

Joint work of Bodlaender, Hans; Cygan, Marek; Kratsch, Stefan; Nederlof, Jesper
Main reference H. Bodlaender, M. Cygan, S. Kratsch, J. Nederlof, "Solving weighted and counting variants of connectivity problems parameterized by treewidth deterministically in single exponential time," arXiv:1211.1505v1 [cs.DS], 2013.
URL http://arxiv.org/abs/1211.1505v1
Main reference H. Bodlaender, M. Cygan, S. Kratsch, J. Nederlof, "Deterministic Single Exponential Time Algorithms for Connectivity Problems Parameterized by Treewidth," in Proc. of the 40th Int'l Colloquium on Automata, Languages, and Programming (ICALP'13), LNCS, Vol. 7965, pp. 196–207, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-39206-1_17

In the talk we discuss properties of two matrices that relate to dynamic programming on tree and path decompositions. The partition matrix has rows and columns indexed by the partitions of t elements; all entries are 1 or 0 depending on whether the meet of the two indexing partitions gives the unit partition. The matching matrix is a submatrix of the partition matrix, obtained by restricting the latter to those partitions that are perfect matchings, i.e., partitions into sets of size two each. We give factorizations for both matrices (and matching lower bounds), proving that the ranks are exactly 2^{t-1} and $2^{t/2-1}$, respectively. This leads to several interesting algorithmic results, among them: 1) The first deterministic algorithms for solving connectivity problems like Hamiltonian Cycle and Steiner Tree in time $O^*(c^{tw})$ when provided a tree decomposition of width tw. 2) A $O((2+\sqrt(2))^{pw})$ time randomized algorithm for Hamiltonian Cycle, given a path decomposition of width pw. The latter runtime is tight under Strong ETH.

5.13 Shortest superstring for short strings

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

Joint work of Golovnev, Alexander; Kulikov, Alexander S.; Mihajlin, Ivan
Main reference A. Golovnev, A. S. Kulikov, I. Mihajlin, "Solving 3-Superstring in $3^{n/3}$ Time," in Proc. of the 38th Int'l Symp. on Mathematical Foundations of Computer Science (MFCS'13), LNCS, Vol. 8087, pp. 480–491, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-40313-2_43

In the shortest common superstring problem (SCS) one is given a set $s_1, \ldots, s_n$ of n strings and the goal is to find a shortest string containing each s_i as a substring. While many approximation algorithms for this problem have been developed, it is still not known whether it can be solved exactly in fewer than 2^n steps. In this paper we present an algorithm that solves the special case when all of the input strings have length 3 in time $3^{n/3}$ and polynomial space. The algorithm generates a combination of a de Bruijn graph and an overlap graph, such that a SCS is then a shortest directed rural postman path (DRPP) on this graph. We show that there exists at least one optimal DRPP satisfying some natural properties. The algorithm works basically by exhaustive search, but on the reduced search space of such paths of size $3^{n/3}$.

5.14 Efficient Computation of Representative Sets and Applications

Daniel Lokshtanov (University of Bergen, NO)

We give computationally efficient variants of "Bolobás' lemma" and its generalization to half-spaces, and consider applications in exact and parameterized algorithms as well as in kernelization. The talk is a mash-up of arxiv:1304.4626, arXiv:1111.2195 as well as a few other results.

5.15 Communication is bounded by root of rank

Shachar Lovett (University of California – San Diego, US)

Main reference S. Lovett, "Communication is bounded by root of rank," arXiv:1306.1877v2 [cs.CC], 2013.
URL http://arxiv.org/abs/1306.1877v2

The log rank conjecture is one of the fundamental open problems in communication complexity. It speculates that if for a boolean $f(x, y)$, its associated matrix $M_{x,y} = f(x, y)$ has rank r (over the reals), then there exists a deterministic protocol computing f which uses only polylog(r) bits of communication. There is a trivial protocol which uses r bits of communication, and further progress on this problem reduced it to $O(r)$ bits [Kotlov-Lovász'96, Kotlov '97] and more recently to $O(r/\log(r))$ bits assuming a number theoretic conjecture [BenSasson-L-RonZewi'12]. In this work, we prove an (unconditional) upper bound of $O(r^{1/2}\log(r))$ bits.

5.16 A subexponential parameterized algorithm for Subset TSP on planar graphs

Dániel Marx (Hungarian Academy of Sciences, HU)

Joint work of Klein, Philip N.; Marx, Dániel

Given a graph G and a subset S of vertices, the SUBSET TSP problem asks for a shortest closed walk in G visiting all vertices of S. The problem can be solved in time $2^k \cdot n^{O(1)}$ using the classical dynamic programming algorithms of Bellman and of Held and Karp, where $k = |S|$ and $n = |V(G)|$. Our main result is showing that the problem can be solved in time $(2^{O(\sqrt{k}\log k)} + W) \cdot n^{O(1)}$ if G is a planar graph with weights that are integers no greater than W. While similar speedups have been observed for various paramterized problems on planar graphs, our result cannot be simply obtained as a consequence of bounding the treewidth of G or invoking bidimensionality theory. Our algorithm consists of two steps: (1) find a locally optimal solution, and (2) use it to guide a dynamic program. The proof of correctness of the algorithm depends on a treewidth bound on a graph obtained by combining an optimal solution with a locally optimal solution.

5.17 On Problems as Hard as CNF-SAT

Jesper Nederlof (Utrecht University, NL)

Joint work of Cygan, Marek ; Dell, Holger; Lokshtanov, Daniel; Marx, Dániel; Nederlof, Jesper; Okamoto, Yoshio; Paturi, Ramamohan; Saurabh, Saket; Wahlstrom, Magnus
Main reference M. Cygan, H. Dell, D. Lokshtanov, D. Marx, J. Nederlof, Y. Okamoto, R. Paturi, S. Saurabh, M. Wahlström, "On Problems as Hard as CNF-SAT," in Proc. of the 2012 IEEE 27th Annual Conf. on Computational Complexity (CCC'12), pp. 74–84, IEEE, 2012.
URL http://dx.doi.org/10.1109/CCC.2012.36

While exhaustive search remains asymptotically the fastest known algorithm for some basic problems, difficult and non-trivial exponential time algorithms have been found for a myriad of problems, including Graph Coloring, Hamiltonian Path, Dominating Set and 3-CNF-Sat. In some instances, improving these algorithms further seems to be out of reach. The CNF-Sat problem is the canonical example of a problem for which the trivial exhaustive search algorithm runs in time $O(2^n)$, where n is the number of variables in the input formula. While there exist non-trivial algorithms for CNF-Sat that run in time $o(2^n)$, no algorithm was able to improve the growth rate 2 to a smaller constant, and hence it is natural to conjecture that 2 is the optimal growth rate. The strong exponential time hypothesis (SETH) by Impagliazzo and Paturi [JCSS 2001] goes a little bit further and asserts that, for every $\epsilon < 1$, there is a (large) integer k such that k-CNF-SAT cannot be computed in time $2^{\epsilon n}$. In this talk, we show that, for every $\epsilon < 1$, the problems Hitting Set, Set Splitting, and NAE-Sat cannot be computed in time $O(2^{\epsilon n})$ unless SETH fails. Here n is the number of elements or variables in the input. For these problems, we actually get an equivalence to SETH in a certain sense. We conjecture that SETH implies a similar statement for Set Cover, and prove that, under this assumption, the fastest known algorithms for Steinter Tree, Connected Vertex Cover, Set Partitioning, and the pseudo-polynomial time algorithm for Subset Sum cannot be significantly improved. Finally, we justify our assumption about the hardness of Set Cover by showing that the parity of the number of set covers cannot be computed in time $O(2^{\epsilon n})$ for any $\epsilon < 1$ unless SETH fails.

5.18 New Algorithms for QBF Satisfiability, and Implications for Circuit Complexity

Rahul Santhanam (University of Edinburgh, GB)

Joint work of Santhanam, Rahul; Williams, Ryan

We study the complexity of the satisfiability problem for quantified Boolean formulas. We show that satisfiability of quantified CNFs of size poly(n) on n variables with at most q quantifier alternations can be solved in time $2^{n-n^{1/q}}$ by zero-error randomized algorithms. This is the first improvement over brute-force search even for $q = 1$. We also show how to achieve non-trivial savings when $q = \omega(\log(n))$. We then draw a connection between algorithms for QBF satisfiability and lower bounds against non-uniform NC^1. We show that NEXP not in NC^1/poly if (i) There are zero-error randomized algorithms solving satisfiability of quantified CNFs of size poly(n) on n variables with at most q quantifier alternations in time $2^{n-n^{\omega_q(1/q)}}$, or (ii) There are zero-error randomized algorithms solving satisfiability of quantified CNFs of size poly(n) on n variables in time $2^{n-\omega(\log(n))}$. Thus even minor improvements of our algorithmic results will yield new circuit lower bounds.

5.19 Better Than Trivial Algorithm for Multicut

Saket Saurabh (The Institute of Mathematical Sciences – Chennai, IN)

Joint work of Saurabh, Saket; Lokshtanov, Daniel; Suchy, Ondrej

There have been several exact algorithms finding a vertex subset of a graph on n vertices satisfying certain properties like acyclicity, bipartiteness in time c^n, where $c < 2$ is a constant. However, there has been very few exact exponential algorithms for cut problems like Multiway Cut or Multicut. In Multicut, we are given an undirected graph $G = (V, E)$ and a family $T = \{(s_i, t_i) \mid s_i, t_i \in V\}$ and objective is to find a minimum size set S such that in $G - S$ there is no path between any pair of vertices in T. In this talk we present an exact algorithm for Multicut running in time 1.987^n. The algorithm is based on branch and bound and utilizes an interesting measure.

5.20 A Satisfiability Algorithm for Sparse Depth Two Threshold Circuits

Stefan Schneider (University of California – San Diego, US)

Joint work of Schneider, Stefan; Impagliazzo, Russell; Paturi, Ramamohan

We give a nontrivial algorithm for the satisfiability problem for threshold circuits of depth two with a linear number of wires which improves over exhaustive search by an exponential factor. We also get an algorithm for 0-1 Integer Linear Programming with a linear number of constraints. The key idea is to reduce the satisfiability problem to the Vector Domination problem, the problem of checking whether there are two vectors in a given collection of vectors such that one dominates the other component-wise. This is joint work with Russell Impagliazzo and Ramamohan Paturi.

5.21 Solving Sparse Instances of Max SAT via Width Reduction and Greedy Restriction

Kazuhisa Seto (The University of Electro-Communications – Tokyo, JP)

Joint work of Seto, Kazuhisa; Tamaki, Suguru; Sakai Takayuki

We present a moderately exponential time polynomial space algorithm for sparse instances of Max SAT. For instances with n variables and cn clauses, our algorithm runs in time $O(2^{(1-\mu_c)n})$, where $\mu_c = \Omega(\frac{1}{c^2 \log^2 c})$. Previously, an exponential space algorithm with $\mu_c = \Omega(\frac{1}{c \log c})$ was shown by Dantsin and Wolpert [SAT 2006] and a polynomial space algorithm with $\mu_c = \Omega(\frac{1}{2^{O(c)}})$ was shown by Kulikov and Kutzkov [CSR 2007]. Our algorithm is based on the combination of two techniques, width reduction of Schuler and greedy restriction of Santhanam. The approach is flexible in the sense that we also obtain algorithms for Max Conj SAT, Max Not-All-Equal SAT and Max Exact-One SAT by slightly modifying the algorithm for Max SAT.

5.22 Time vs Approximation Trade-offs and Strong Relaxations

David Steurer (Cornell University, US)

I will survey recent results about approximability in the regime of intermediate complexity – strictly between polynomial and exponential time. A unifying theme are strong relaxations, especially the semidefinite programs obtained from sum-of-squares methods.

5.23 A Satisfiability Algorithm and Average-Case Hardness for Formulas over the Full Binary Basis

Suguru Tamaki (Kyoto University, JP)

Joint work of Seto, Kazuhisa; Suguru Tamaki
Main reference K. Seto, S. Tamaki, "A Satisfiability Algorithm and Average-Case Hardness for Formulas over the Full Binary Basis," Computational Complexity, 22(2):245–274, 2013.
URL http://dx.doi.org/10.1007/s00037-013-0067-7

We present a moderately exponential time algorithm for the satisfiability of Boolean formulas over the full binary basis. For formulas of size at most cn, our algorithm runs in time $2^{(1-\mu_c)n}$ for some constant $\mu_c > 0$. As a byproduct of the running time analysis of our algorithm, we obtain strong average-case hardness of affine extractors for linear- sized formulas over the full binary basis.

5.24 SETH and some natural graph problems in poly-time

Virginia Vassilevska Williams (Stanford University, US)

Joint work of Roditty, Liam; Vassilevska Williams, Virginia
Main reference L. Roditty, V. Vassilevska Williams, "Fast approximation algorithms for the diameter and radius of sparse graphs," in Proc. of the 45th Annual ACM Symp. on Theory of Computing (STOC'13), pp. 515–524, ACM, 2013.
URL http://dx.doi.org/10.1145/2488608.2488673

We discuss several reductions from CNF-SAT to some natural graph problems that relate the complexity of these problems to the Strong Exponential Time Hypothesis. For instance, we show that unless SAT is in $(2-\epsilon)^n$ time for some $\epsilon > 0$, any $(3/2-\delta)$-approximation algorithm for the graph diameter must take $\Omega(n^{2-\epsilon})$ time for all $\epsilon > 0$. There is an $\tilde{O}(m\sqrt{n})$-time 3/2-approximation for the problem, and our reduction shows that this algorithm is in a sense optimal, assuming Strong ETH. We also show that some dynamic graph problems admit lower bounds based on Strong ETH. For instance, maintaining the number of strongly connected components of a graph under edge insertions or deletions requires amortized update time $\Omega(m^{1-\epsilon})$ for all $\epsilon > 0$, thus showing that the trivial update algorithm may be optimal.

5.25 Abusing the Tutte matrix: A polynomial compression for the Steiner Cycle problem

Magnus Wahlstroem (Royal Holloway University of London, GB)

Main reference M. Wahlström, "Abusing the Tutte Matrix: An Algebraic Instance Compression for the K-set-cycle Problem," in Proc. of the 30th Int'l Symp. on Theoretical Aspects of Computer Science (STACS'13), LIPIcs, Vol. 20, pp. 341–352, Schlos Dagstuhl – Leibniz-Zentrum für Informatik, 2013.
URL http://dx.doi.org/10.4230/LIPIcs.STACS.2013.341

We give an algebraic, determinant-based algorithm for the Steiner Cycle problem, i.e., the problem of finding a cycle through a set K of specified elements. Our approach gives a simple FPT algorithm for the problem, matching the $O^*(2^{|K|})$ running time of the algorithm of Björklund et al. (SODA, 2012). Furthermore, our approach is open for treatment by classical algebraic tools (e.g., Gaussian elimination), and we show that it leads to a polynomial compression of the problem, i.e., a polynomial-time reduction of the Steiner Cycle problem into an algebraic problem with coding size $O(|K|^3)$. This is surprising, as several related problems (e.g., k-Cycle and the Disjoint Paths problem) are known not to admit such a reduction unless the polynomial hierarchy collapses. Furthermore, despite the result, we are not aware of any witness for the Steiner Cycle problem of size polynomial in $|K| + \log n$, which seems (for now) to separate the notions of polynomial compression and polynomial kernelization (as a polynomial kernelization for a problem in NP necessarily implies a small witness).

5.26 The Organizers Try to Explain ETH Using Somebody Else's Slides

Ryan Williams (Stanford University, US)

To begin the seminar with a common focal point, I presented an overview of the Exponential Time Hypothesis (ETH) and the Strong Exponential Time Hypothesis (SETH). Namely, I discussed how these hypotheses are formulated, their consistency with current knowledge, their many consequences for parameterized and exact exponential time algorithms, and (briefly) their plausibility. For this purpose, I worked with a modification of excellent slides by Daniel Lokshtanov (hence the talk title), based on the survey [1].

References

1 Daniel Lokshtanov, Dániel Marx, Saket Saurabh. Lower bounds based on the Exponential Time Hypothesis. *Bulletin of the EATCS* 105, 41–72, 2011.

6 Open Problems

6.1 Fast algorithm for DRPP w.r.t. the number of weakly connected components

Alexander S. Kulikov

In the *directed rural postman problem* (DRPP) one is given a directed weighted multigraph $G = (V, A)$ together with a subset $R \subseteq A$ of its arcs and the goal is to find a shortest closed walk in this graph going through all the arcs from R. Although DRPP is NP-hard in general case, it can be solved in polynomial time if the arcs from R form a single weakly connected component [1] (weakly connected components of a directed graph are connected components in this graph with all directed arcs replaced by undirected edges).

Question

Is DRPP fixed-parameter tractable w.r.t. the number of weakly connected components k? Can it be solved in time $O^*(2^k)$?

Remark. Shortly after the seminar Gutin, Wahlström, and Yeo [2] answered the first question in the affirmative. One month later they answered the second question as well.

References

1 N. Christofides, V. Campos, A. Corberan, and E. Mota. An algorithm for the Rural Postman problem on a directed graph. In *Netflow at Pisa*, volume 26 of *Mathematical Programming Studies*, pages 155–166. Springer Berlin Heidelberg, 1986.
2 G. Gutin, M. Wahlström, and A. Yeo.Parameterized rural postman and conjoining bipartite matching problems. *arXiv preprint arXiv:1308.2599*, 2013.

6.2 Shortest common superstring problem

Alexander S. Kulikov

In the *shortest common superstring problem* (SCS) one is given n strings $s_1, \ldots, s_n$ and the goal is to find a shortest string containing each s_i as a substring. It can be solved in $O^*(2^n)$ time and space by dynamic programming. Using inclusion-exclusion one can reduce the space to polynomial.

Question

Can SCS be solved in time $O^*(1.99^n)$? Or, at least, can it be solved in time $O^*((2-c)^n)$ where $c = c(r) > 0$ for a special case when all input strings have length at most r? Can one reduce TSP with n vertices to SCS with n strings (this would show that SCS is as hard as TSP)?

Remark. Note also that the well-known greedy conjecture (saying that one gets a 2-approximation for SCS by overlapping $n-1$ times two strings with maximal overlap) is still open.

6.3 The exponential complexity of some SUBSET SUM variants

Andrew Drucker

Question

Here are some variants of the NP-complete SUBSET SUM problem.

In the THRESHOLD PROBABILITY problem, one is given integers $a_1, \ldots, a_n$ and a threshold value T, and asked to compute the probability over a random $S \subseteq [n]$ that $\sum_{i \in S} a_i \geq T$. This problem lies in $\#P$.

In the RANKED SUBSET SUM problem, one is given integers $a_1, \ldots, a_n$ and a number $k \in \{1, 2, \ldots, 2^n\}$, and is asked to compute the k^{th} largest value $\sum_{i \in S} a_i \geq T$, ranging over all possible $S \subseteq [n]$. It is not clear to me whether this problem is even in $\#P$. It does lie in $P^{\#P}$, however. (Proof sketch: we use binary search. With a $\#P$ oracle we can determine the rank order of a given subset S's sum. Using the power of the Polynomial Hierarchy we can choose a subset nearly-uniformly from those ranked above or below S. And computations in the Polynomial Hierarchy can be done in $P^{\#P}$ by Toda's theorem.)

It may well be known that each of these problems is $\#P$-hard. However, my question is whether a stronger statement is true: namely, whether either or both of these problems requires $2^{\Omega(n)}$ time (or something close) under the *counting Exponential Time Hypothesis (#ETH)* of Dell, Husfelt, Marx, and Taslaman [1]. These would give more basic examples of problems to which this hardness assumption applies.

References

1 Holger Dell and Thore Husfeldt and Dániel Marx and Nina Taslaman and Martin Wahlen, *Exponential Time Complexity of the Permanent and the Tutte Polynomial*, CoRR, abs/1206.1775, 2012, http://arxiv.org/abs/1206.1775.

6.4 Maximum Acyclic Subgraph

Eunjung Kim

Given a directed graph $D = (V, A)$, the problem Maximum Acyclic Subgraph, or MAS, asks to find a subdigraph $D' = (V, A')$ of D such that $|A'|$ is maximized. The problem MAS can be solved in time $O(2^n)$, where $n = |V|$, using dynamic programming [1].

Question

Is there an algorithm for MAS running in time $O(c^n)$ for a constant $c < 2$?

Remark. The problem MAS can be viewed as one of the family of arity k ordering CSP. An instance of ordering CSP consists of a set of variables V and a set of constraints $\mathcal{C}$, in which constraints are tuples of elements of V. The goal is to find a total ordering of the variables, $\pi : V \to \{1, \ldots, |V|\}$, which satisfies as many constraints as possible. A constraint $(v_1, \ldots, v_k)$ is satisfied by an ordering π when $\pi(v_1) < \pi(v_2) < \ldots < \pi(v_k)$. An instance has arity k if all the constraints involve at most k elements. The problem MAS is equivalent to arity 2 ordering CSP.

While the dynamic programming approach for ordering problem such as in [1] yields an $O(2^n)$-time algorithm for arity 3 ordering CSP, it is unlikely to solve arity k ordering CSP in time $O(2^{o(n \log n)})$ for $k \geq 4$ under ETH, see [2]. Such dichotomy leads to a further question: can arity 3 ordering CSP be solved in time $O(c^n)$ for $c < 2$?

References

1 Hans L. Bodlaender, Fedor V. Fomin, Arie M. C. A. Koster, Dieter Kratsch, and Dimitrios M. Thilikos. A note on exact algorithms for vertex ordering problems on graphs. *Theory Comput. Syst.*, 50(3):420–432, 2012.

2 Eun Jung Kim and Daniel Gonçalves. On exact algorithms for the permutation CSP. *Theoretical Computer Science*, Vol. 514, pp. 109–116, 2012. ISSN 0304-3975. http://dx.doi.org/10.1016/j.tcs.2012.10.035.

6.5 Enumerating all story arcsets (SAS) of a digraph with polynomial delay

Ewald Speckenmeyer

The problem was motivated by a biological question w.r.t. metabolic networks.
$G = (V, E)$ be a digraph. A subset $F \subseteq E$ of arcs is a **feedback arc set** (**FAS**) if $G - F$ is acydic.

Result

Schwikowski, Sp.: On enumerating all minimal solutions of feedback problems; DAM 2002

All minimal FAS of a digraph can be enumerated with polynomial delay.
$\rightsquigarrow$ All maximal arc-induced acydic digraphs of a digraph can be enumerated with polynomial delay.

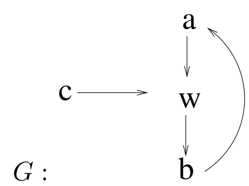
Nodeset $V = B \cup W$ consists of **black** nodes from B and **white** nodes from W. Definitions used here are from:

- Acuña, et al.: Telling stories; TCS 2012
- Borassi, et al: Telling stories fast; SEA 2013.

A **pitch** of $G = (B \cup W, E)$ is an acydic digraph G' of G, containing all nodes from B, a subset $W' \subseteq W$, and a subset $E' \subseteq E$, s.t. no node from W' is source or target node of G'. A **story** S of G is a maximal pitch of G. The corresponding arc set $E - E'$ is called **SAS** of G. **Telling stories** means enumerating all stories S of G.

Example

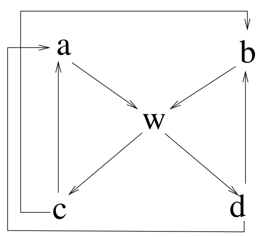
$B = \{a, b, c\}, W = \{w\}$



3 minimal FAS: $\{a \to w\}$, $\{\mathbf{w} \to \mathbf{b}\}$ $\{b \to a\}$. $E(G) - \{\mathbf{w} \to \mathbf{b}\}$ is no SAS. An SAS induces a FAS, but not vice versa!

Example

(from Acuña, et al.)



$S = \{a \to w,\ b \to w,\ w \to c,\ w \to d\}$ is a SAS, but $E - S$ is **not** a minimal FAS of G.

Open Problem

Enumerate all stories of a digraph with polynimial delay. A first step towards solving the problem by Borassi, et.al.: Enumerate pitches with linear delay.

6.6 Exact Algorithm For Subgraph Isomorphism

Fedor V. Fomin

For any graph G with n vertices and m edges, it is possible to decide in time $2^{m+o(n)}$ if G contains a given graph F as a subgraph: try all possible edge subsets of G and for each subset check if the obtained graph is isomorphic to F. Another approach is to try all the permutations of the vertices of G and F, and for each of these permutations, to compare vertex neighborhoods. This will give us running time $O(n!n^2) = 2^{O(n \log n)}$.

Question

Can subgraph isomorphism can be solved in time $2^{o(n \log n)}$?

6.7 Exact algorithms for Special Treewidth and Spaghetti Treewidth

Hans Bodlaender

Courcelle [2] introduced the notion of *Special Treewidth.* Some resuls (e.g., fixed parameter tractability) on the notion were recently obtained [1]. A *special tree decomposition* is a rooted tree decomposition, such that for each vertex, the bags containing the vertex form a rooted path in the tree from the tree decomposition. The special treewidth is the minimum width of a special tree decomposition. (A rooted path is a path from a vertex to one of its ancestors.)

In [1], we also introduced the notion of spaghetti treewidth. A *spaghetti tree decomposition* is a tree decompositions, such that for each vertex, the bags containing the vertex from a pah in the tree. (I.e., it is not necessarily rooted.)

Computing the special treewidth and spaghetti treewidth are NP-hard.

Question

Find efficient exact algorithms to compute the special treewidth or spaghetti treewidth of a given graph.

Remark. The existence of $O*(2^n)$ algorithms for these problems is not known.

References

1 Hans L. Bodlaender, Vincent Kreuzen and Stefan Kratsch. *Fixed-Parameter Tractability and Characterizations of Small Special Treewidth* , To appear in the proceedings of WG 2013.

2 Bruno Courcelle. Special tree-width and the verification of monadic second-order graph properties. Proceedings FSTTCS 2010, LIPIcs, Vol. 8, pp. 13–29, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, DOI: 10.4230/LIPIcs.FSTTCS.2010.13, 2010.

6.8 Treewidth of Planar Graphs

Hans Bodlaender

The famous Ratcatcher algorithm by Seymour and Thomas gives a polynomial time algorithm to compute the branchwidth of a planar graph. This directly implies that the treewidth of a planar graph can be approximated with a factor 1.5. The treewidth of a planar graph with n vertices is bounded by $O(\sqrt{n})$.

Question

How hard is the following problem: Given is a planar graph G, and an integer k. Is the treewidth of G at most k? Is it NP-complete? Is it polynomial time solvable?

Comment. This is a long standing open problem.

References

1 aul D. Seymour, Robin Thomas: Call Routing and the Ratcatcher. Combinatorica 14(2): 217-241 (1994)

6.9 Improving pseudo polynomial time algorithms for Subset Sum

Jesper Nederlof

The Subset Sum problem is defined as follows: Given integers $w_1, \ldots, w_n$ and budget w, decide whether there exists $X \subseteq \{ 1, \ldots, n\}$ such that $\sum_{i \in X} w_i = w$. A classical example of an exact algorithm for a NP-hard problem is the algorithm of Bellman from the 50's that solves the Subset Sum problem in $O(nw)$ time and $O(w)$ space (note that this resource bound assumes that computation and storage involving arbitrary integers requires constant time / storage).

Question 1

Does there exist $\epsilon > 0$ such that Subset Sum be solved in $w^{1-\epsilon}n^{O(1)}$?

In [1], the authors relate the above question to improvements of naive algorithms for Set Cover. However, there are no known consequences of the Strong ETH in this context.

In 2010 [4] (see also [2, 3]), alternatives for Bellman's algorithm were presented that are significantly more space efficient. In particular, Subset Sum can be solved in $\tilde{O}(n^3 t \log t)$ time and $\tilde{O}(n^2)$ space (note that these resource bounds are stated in terms of bit operations / storage). It would be interesting to see whether the time usage of this algorithm can be improved in order to practically compete/outperform Bellman's algorithm:

Question 2

Can Subset Sum be solved in $\tilde{O}(nt \log t)$ time and $\tilde{O}(n)$ space?

After posing this question, Ryan Williams mentioned that, together with Huacheng Yu, he made partial progress in this direction.

In the Knapsack problem we are additionally given values $v_1, \dots, v_n$, goal v and we are asked to find $X \subseteq \{1, \dots, n\}$ such that $\sum_{i \in X} v_i \geq v$ and $\sum_{i \in X} w_i \leq w$. Bellman's approach solves the Knapsack problem in $O(n \min\{v, w\})$ and $O(\min\{v, w\})$ space. However, considering space efficient algorithms, Knapsack can currently only be solved in $\tilde{O}(n^4 vw \log(vw))$ time and $\tilde{O}(n^2 \log(vw))$ space. It would be interesting to improve this:

Question 3

Can Knapsack be solved in $\min\{v, w\} n^{O(1)}$ time and $(n \log(vw))^{O(1)}$ space?

Other non-trivial improvements or time/space tradeoffs in this context would also be interesting, but it seems hard to mix the greedy aspect of Bellman's approach with the coefficient interpolation strategy of the space efficient approach.

References

1 Marek Cygan, Holger Dell, Daniel Lokshtanov, Dániel Marx, Jesper Nederlof, Yoshio Okamoto, Ramamohan Paturi, Saket Saurabh, and Magnus Wahlström. On Problems as Hard as CNF-sat. In *IEEE Conference on Computational Complexity*, pages 74–84. IEEE, 2012. ISBN 978-1-4673-1663-7.

2 Michael Elberfeld, Andreas Jakoby, and Till Tantau. Logspace Versions of the Theorems of Bodlaender and Courcelle. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 143–152, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4244-7. 10.1109/FOCS.2010.21. URL http://dx.doi.org/10.1109/FOCS.2010.21.

3 Daniel M. Kane. Unary Subset-Sum is in Logspace. *CoRR*, abs/1012.1336, 2010.

4 Daniel Lokshtanov and Jesper Nederlof. Saving Space by Algebraization. In *Proceedings of the 42nd ACM symposium on Theory of computing*, STOC '10, pages 321–330, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0050-6. 10.1145/1806689.1806735. URL http://doi.acm.org/10.1145/1806689.1806735.

6.10 Channel assignment

Łukasz Kowalik

In the channel assignment problem, we are given a symmetric weight function $w : V^2 \to \mathbb{N}$ (we assume that $0 \in \mathbb{N}$). The elements of V are called vertices (as w induces a graph on the vertex set V with edges corresponding to positive values of w). We say that w is ℓ-bounded when for every $x, y \in V$ we have $w(x, y) \leq \ell$. An assignment $c : V \to \{1, \dots, s\}$ is called *proper* when for each pair of vertices x, y we have $|c(x) - c(y)| \geq w(x, y)$. The number s is called the *span* of c. The goal is to find a proper assignment of minimum span. Note that the special case when w is 1-bounded corresponds to the classical graph coloring problem.

Currently the fastest algorithm for the channel assignment problem, due to Cygan and Kowalik [1] runs in $O^*((\ell + 1)^n)$ time, where n denotes the number of vertices.

Traxler [2] has shown that for any constant c, the Constraint Satisfaction Problem (CSP) has no $O(c^n)$-time algorithm, assuming the Exponential Time Hypothesis (ETH). More precisely, he shows that ETH implies that CSP requires $d^{\Omega(n)}$ time, where d is the domain size. On the other hand, graph coloring, which is a variant of CSP with unbounded domain, admits a $O^*(2^n)$-time algorithm. The channel assignment problem is a generalization of graph coloring and a special case of CSP. In that context, the central open problem in the complexity of the channel assignment problem is the following:

Open Problem

Find an $O^*(c^n)$-time algorithm for a constant c independent of ℓ or to show that such the algorithm does not exist, assuming ETH (or other well-established complexity conjecture).

References

1 M. Cygan and Ł. Kowalik. Channel Assignment via Fast Zeta Transform. *Information Processing Letters*, 111(15):727-730, 2011.

2 P. Traxler. The Time Complexity of Constraint Satisfaction. In *Proc. IWPEC 2008*, LNCS 5018, pages 190–201, 2008.

6.11 Improve exponential time algorithm for Bandwidth

Marek Cygan

In the Bandwidth problem one is given an undirected graph $G = (V, E)$ together with an integer b. The question is whether there exists a bijection $\pi : V \to \{1, \ldots, n\}$, such that for each edge $uv \in E$ we have $|\pi(u) - \pi(v)| \leq b$. Bandwidth is an NP-complete problem, and currently the fastest known exact exponential time algorithms work in:

- $O(4.383^n)$ time and exponential space [1],
- $O(9.363^n)$ time and polynomial space [2].

Question

Can the above algorithms be improved, in particular is there an $O^*(4^n)$ time algorithm for Bandwidth?

References

1 Marek Cygan and Marcin Pilipczuk, *Exact and approximate bandwidth*, Theoretical Computer Science 411(40-42): 3701-3713 (2010).

2 Marek Cygan and Marcin Pilipczuk, *Bandwidth and distortion revisited*, Discrete Applied Mathematics, 160(4-5): 494-504 (2012).

6.12 Relation between permanent computation and SETH

Marek Cygan

One can compute permanent of a given matrix in $O^*(2^n)$ time either by dynamic programming or by the inclusion-exclusion principle. A slightly faster $2^{n-\Omega(\sqrt{n/\log n})}$ was recently obtained by Björklund [1]. Can we relate the hardness of permanent computation to SETH, by showing that one of those problems is at least as hard as the other?

Remark. Preferably the reduction would involve binary matrices only, but the general case would also be interesting.

References

1 Andreas Björklund, *Below All Subsets for Permutational Counting Problems*, CoRR, abs/1211.0391, 2012, http://arxiv.org/abs/1211.0391.

6.13 Exact Counting of Linear Extensions

Mikko Koivisto

Counting the linear extensions of a given partially ordered set (poset) is #P-complete [1]. While the problem admits a FPRAS [4], the practical value of the existing schemes [2] is questionable, as the running time grows roughly as $n^6\epsilon^{-2}$ for n-element posets and relative error ϵ. Fast exact algorithms could be more practical when n is around 100 and good accuracy is needed. Currently the best known worst-case bound is $2^n n^{O(1)}$, based on simple dynamic programming across the downsets of the poset [3, 5].

Question

Is there a $O(c^n)$-time algorithm for counting linear extensions, for some $c < 2$? Is the problem fixed-parameter tractable with respect to the treewidth of the Hasse diagram (i.e., the transitive reduction of the poset)?

Remark. The problem can be solved in $n^{t+O(1)}$ time for treewidth t (unpublished work).

References

1 G. Brightwell and P. Wrinkler, *Counting linear extensions*, Order 8 (1991) 225–242.
2 R. Bubley and M. Dyer, *Faster random generation of linear extensions*, Discrete Mathematics 201 (1999) 81–88.
3 K. De Loof, H. De Meyer, and B. De Baets, *Exploiting the lattice of ideals representation of a poset*, Fundamenta Informaticae 71 (2006) 309–321.
4 M. Dyer, A. Frieze, and R. Kannan, *A random polynomial-time algorithm for approximating the volume of convex bodies*, Journal of the ACM 38 (1991) 1–17.
5 T. Niinimäki and M. Koivisto, *Annealed importance sampling for structure learning in Bayesian networks*, 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013), pp. 1579–1585, AAAI, 2013.

6.14 Vector Positive Explanation

Rolf Niedermeier

The problem Vector Postitive Explanation is motivated by applications in context of radiation therapy.

Input: A vector $A \in I\!N_0^n$ and an integer $k > 0$.
Question: Can A be explained by at most k positive segments?

Herein, a *segment* is a vector in $\{0, a\}^n$ for some $a \in \mathbb{Z} \setminus \{0\}$ where all a-entries occur consecutively, and a segment is positive if a is positive. An *explanation* is a set of segments that component-wisely sum up to the input vector.

The parameter we are interested in is the maximum difference between two consecutive vector entries $\delta := \max_{0 \leq i \leq n} |A[i] - A[i+1]|$, assuming that $A[0] = A[n+1] = 0$.

Example. The vector $(4, 3, 3, 4)$ can be explained by $(3, 3, 3, 3)$, $(1, 0, 0, 0)$, and $(0, 0, 0, 1)$. Herein, $\delta = 4$.

Question

Is Vector Positive Explanation parameterized by δ fixed-parameter tractable?

Remark. The variant where also negative segments are allowed is known to be fixed-parameter tractable when parameterized by δ (using an integer linear program formulation and applying Lenstra's famous result).

References

1 Robert Bredereck and Jiehua Chen and Sepp Hartung and Christian Komusiewicz and Rolf Niedermeier and Ondrej Suchý, *On Explaining Integer Vectors by Few Homogenous Segments*, Proc. WADS 2013, LNCS 8037, pages 207-218, 2013. Journal version under review and available upon request.

6.15 The Complexity of $\#k$-SAT

Rahul Santhanam

Question

Is there a deterministic algorithm for $\#k$-SAT running in time $2^{n-\Omega(n/k)}$?

Remarks. There is a probabilistic algorithm due to Impagliazzo, Mathews and Paturi with the required time complexity, but it is not known how to derandomize the algorithm. To the best of my knowledge, it is also unknown whether there is an algorithm for k-SAT running in time $2^{n-n/k}$.

6.16 Strong Backdoor detection

Serge Gaspers

Let $\mathcal{C}$ be a class of CNF formulas. For example, the class $\mathcal{W}_t$ contains all CNF formulas whose incidence graphs have treewidth at most t. The *incidence graph* of a CNF formula is a bipartite graph whose vertex set is bipartitioned into the set of variables and the set of clauses of F; a variable is adjacent to a clause if the variable is contained in the clause.

The Strong $\mathcal{C}$-Backdoor Detection problem [3] takes as input a CNF formula F on n variables and an integer k, and the question is whether there exists a subset of variables $S \subseteq \mathsf{var}(F)$ with $|S| \leq k$ such that for each assignment $\alpha : S \to \{0, 1\}$ we have that $F[\alpha] \in \mathcal{C}$. Here, the formula $F[\alpha]$ is obtained from F by removing all clauses containing a literal set to 1 by α, and by removing the literals set to 0 from all remaining clauses.

Question 1

Let $t > 0$. Is there an algorithm for Strong $\mathcal{W}_t$-Backdoor Detection with running time $O((3-\epsilon)^n)$ for some $\epsilon > 0$?

Note that a $O(3^n)$ time algorithm is trivially obtained by going through all $S \subseteq \mathsf{var}(F)$ and all assignments to S. Since backdoors are mainly useful when they are small and since the main reason for computing backdoors is to solve SAT and #SAT, we arrive at the second question.

Question 2

Let $t > 0$. Is there an algorithm for Strong $\mathcal{W}_t$-Backdoor Detection restricted to instances where $k \le n/4$ with running time $O((2-\epsilon)^n)$ for some $\epsilon > 0$?

Note that the trivial algorithm restricted to subsets of size at most $n/4$ executes at least $\binom{n}{n/4}2^{n/4} \approx 2.0868\ldots^n$ steps. Both questions remain open and interesting for $t = 1$. For further background on these problems, see [2, 1].

References

1. Serge Gaspers and Stefan Szeider, *Strong Backdoors to Bounded Treewidth SAT*, Proceedings of FOCS 2013, 2013.
2. Serge Gaspers and Stefan Szeider, *Backdoors to Satisfaction*, The Multivariate Algorithmic Revolution and Beyond, pages 287-317, 2012.
3. Ryan Williams, Carla P. Gomes, and Bart Selman, *Backdoors To Typical Case Complexity*, Proceedings of IJCAI 2003, pages 1173-1178, 2003.

6.17 Monochromatic Rectangles in Low-Rank Matrices

Shachar Lovett

Conjecture: Let A be a real $n \times n$ matrix of rank r, such that (say) $n^2/2$ of its entries are zero. Then there exists a rectangle $R \subset [n] \times [n]$ of size $|R| \ge n^2 \exp(-O(\sqrt{r}))$ such that A restricted to R is zero.

If true, the bound is best possible. Take $A = BB^t$ where B is an $n \times r$ matrix whose rows are all the $\{0,1\}^r$ vectors of hamming weight $\sqrt{r}/10$. We have $n = r^{\Theta(\sqrt{r})}$. Then 99% of the elements of A are zero, but the largest zero rectangle corresponds to choosing rows corresponding to vectors supported on the first half of the coordinates, and columns corresponding to vectors supported on the last half of the coordinates.

6.18 Cutwidth

Saket Saurabh

Question

The cutwidth of a graph G is the smallest integer k such that the vertices of G can be arranged in a linear layout $[v_1, \ldots, v_n]$ in such a way that, for every $i = 1, \ldots, n-1$, there are at most k edges with one endpoint in $\{v_1, \ldots, v_i\}$ and the other in $\{v_{i+1}, \ldots, v_n\}$. Does there exist an algorithm to find cutwidth of a graph on n vertices that runs in time c^n, where $c < 2$ is a constant.

Remark. There does exist an algorithm on bipartite graphs that run in time 1.415^n [1].

References

1 M. Cygan, D. Lokshtanov, M. Pilipczuck, M. Pilipczuck and S. Saurabh, *On Cutwidth Parameterized by Vertex Cover*, IPEC, LNCS 7112, 246-258, 2011.

6.19 Converting CNF to DNF

Stefan Schneider

Question

Given a k-CNF, how large is the smallest equivalent DNF, measured in the number of terms?

This question is discussed by Miltersen, Radhakrishnan and Wegener [1], who show that $2^{n(1-\frac{1}{100k})}$ terms are sufficient. The bound coincides up to the constant in the exponent with block parity, that requires $2^{n(1-\frac{1}{k})}$ terms (if k divides n). The question asks if we can close the gap between the lower bound and the upper bound.

The proof of Miltersen, Radhakrishnan and Wegener is based on Håstad's Switching Lemma [2]. Since the Switching Lemma is not tight up to constants, the resulting bound on the number of terms has this constant in the exponent.

To close this gap, one might try to prove a bound based on the Satisfiability Coding Lemma [3], which shows (among others) that no k-CNF can have more than $2^{n(1-\frac{1}{k})}$ isolated solutions, where an isolated solution is a satisfying assignment such that no other assignment of Hamming distance 1 is also satisfying.

References

1 P. Miltersen, J. Radhakrishnan and I. Wegener , *On converting CNF to DNF*, Theor. Comput. Sci, 2005.
2 J. Håstad, *Almost optimal lower bounds for small depth circuits*, STOC, 1986.
3 R. Paturi, P. Pudlák and F. Zane, *Satisfiability Coding Lemma*, FOCS, 1997.

6.20 Directed Hamiltonicity

Thore Husfeldt

Despite recent progress in exponential time algorithms for Hamiltonicity [2, 3, 5, 4], Open Problem 3.1 in Woeginger's influential survey [7] remains open in its general form: we do not have an algorithm for Hamilonicity in directed graphs (nor for TSP with unbounded weights) whose exponent is better than the dynamic programming solution of Bellman and Held–Karp from 1962 [1, 6].

Question

Construct an algorithm for the Hamiltonian cycle problem in directed graphs with time complexity $O(1.999^n)$.

References

1 R. Bellman. Dynamic programming treatment of the travelling salesman problem, J. Assoc. Comput. Mach. 9, pp. 61–63, 1962.
2 A. Björklund. Determinant sums for undirected Hamiltonicity. In Proceedings of the 51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23–26, 2010, Las Vegas, Nevada, USA. IEEE Computer Society 2010, 51st FOCS, pages 173–182.
3 A. Björklund. Below all subsets for permutational counting problems. arXiv:1211.0391, 2012.
4 A. Björklund, T. Husfeldt: The Parity of Directed Hamiltonian Cycles. 54th Annual Symposium on Foundations of Computer Science, FOCS 2013, October 27–29, 2013, Berkeley, California, arXiv:1301.7250, 2013.
5 M. Cygan, S. Kratsch, and J. Nederlof. Fast Hamiltonicity checking via bases of perfect matchings. 45th ACM Symposium on the Theory of Computing, STOC 2013, June 1–4, 2013, Palo Alto, USA, arXiv:1211.1506, 2012.
6 M. Held and R. M. Karp. A dynamic programming approach to sequencing problems, J. Soc. Indust. Appl. Math. 10, pp. 196–210, 1962.
7 G. J. Woeginger. Exact algorithms for NP-hard problems: A survey. In Combinatorial Optimization – Eureka! You shrink!, M. Juenger, G. Reinelt and G. Rinaldi (eds.). Lecture Notes in Computer Science 2570, Springer 2003, pages 185–207.

Participants

- Per Austrin
KTH Stockholm, SE
- Hans L. Bodlaender
Utrecht University, NL
- Marek Cygan
University of Warsaw, PL
- Holger Dell
University of Wisconsin – Madison, US
- Andrew Drucker
MIT, US
- Michael Elberfeld
ICSI – Berkeley, US
- Fedor V. Fomin
University of Bergen, NO
- Serge Gaspers
The University of New South Wales, AU
- Pinar Heggernes
University of Bergen, NO
- Timon Hertli
ETH Zürich, CH
- Thore Husfeldt
IT Univ. of Copenhagen, DK
- Kazuo Iwama
Kyoto University, JP
- Petteri Kaski
Aalto University, FI
- Eun Jung Kim
University Paris-Dauphine, FR
- Joachim Kneis
RWTH Aachen, DE
- Mikko Koivisto
University of Helsinki, FI
- Łukasz Kowalik
University of Warsaw, PL
- Dieter Kratsch
University of Metz, FR
- Stefan Kratsch
TU Berlin, DE
- Alexander S. Kulikov
Steklov Institute – St. Petersburg, RU
- Daniel Lokshtanov
University of Bergen, NO
- Shachar Lovett
University of California – San Diego, US
- Dániel Marx
Hungarian Acad. of Sciences, HU
- Matthias Mnich
Universität des Saarlandes, DE
- Jesper Nederlof
Utrecht University, NL
- Rolf Niedermeier
TU Berlin, DE
- Ramamohan Paturi
University of California – San Diego, US
- Peter Rossmanith
RWTH Aachen, DE
- Rahul Santhanam
University of Edinburgh, GB
- Saket Saurabh
The Institute of Mathematical Sciences – Chennai, IN
- Dominik Scheder
Aarhus University, DK
- Stefan Schneider
University of California – San Diego, US
- Kazuhisa Seto
The Univ. of Electro-Communications – Tokyo, JP
- Gregory B. Sorkin
London School of Economics, GB
- Ewald Speckenmeyer
Universität Köln, DE
- David Steurer
Cornell University, US
- Stefan Szeider
TU Wien, AT
- Navid Talebanfard
Aarhus University, DK
- Suguru Tamaki
Kyoto University, JP
- Virginia Vassilevska Williams
Stanford University, US
- Magnus Wahlström
Royal Holloway University of London, GB
- Ryan Williams
Stanford University, US

Report from Dagstuhl Seminar 13341

Verifiably Secure Process-Aware Information Systems

Edited by

Rafael Accorsi[1], Jason Crampton[2], Michael Huth[3], and Stefanie Rinderle-Ma[4]

1 **Universität Freiburg, DE,** rafael.accorsi@iig.uni-freiburg.de
2 **RHUL – London, GB,** jason.crampton@rhul.ac.uk
3 **Imperial College London, GB,** M.Huth@imperial.ac.uk
4 **Universität Wien, AT**

Abstract

From August 18–23, 2013, the Dagstuhl Seminar "Verifiably Secure Process-aware Information Systems" was held in Schloss Dagstuhl – Leibniz Center for Informatics. During this seminar, participants presented their current research and discussed open problems in the arising field of securing information systems driven by processes. The executive summary and abstracts of the talks given during the seminar are put together in this paper.

Seminar 18.–23. August, 2013 – www.dagstuhl.de/13341

1998 ACM Subject Classification C.2.2 Protocol Verification, C.4 Reliability, availability, and serviceability, D.2.0 Protection Mechanisms, D.2.4 Software/Program Verification, D.2.11 Software Architectures, D.4.5 Reliability, D.4.6 Security and Protection, F.3.2 Process models, H.4.1 Workflow management, I.2.8 Scheduling, J.1 Administrative Data Processing, K.5.2 Regulation

Keywords and phrases Business Processes, Information Security, Compliance, Risk-Aware Processes, Service Compositions

Digital Object Identifier 10.4230/DagRep.3.8.73

1 Executive Summary

Rafael Accorsi
Jason Crampton
Michael Huth
Stefanie Rinderle-Ma

Business processes play a major role in many commercial software systems and are of considerable interest to the research communities in Software Engineering, and Information and System Security. A process-aware information system provides support for the specification, execution, monitoring and auditing of intra- as well as cross-organizational business processes.

Designing and enacting secure business processes is as tricky as "Programming Satan's Computer", as Ross Anderson and Roger Needham observed in a paper with that title. Recent fraud disasters show how subtle secure process engineering and control can be. The Swiss bank UBS suffered from a rogue trader scandal in 2011, which led to a loss of a then-estimated US$2 billion, was possible because the risk of trades could be disguised by using "forward-settling" Exchange-traded Funds (ETF) cash positions. Specifically, processes that implemented ETF transactions in Europe do not issue confirmations until after settlement has taken place. The exploitation of this process allows a party in a transaction to receive payment for a trade before the transaction has been confirmed. While the cash proceeds in this scheme cannot be simply retrieved, the seller may still show the cash on their books and possibly use it in further transactions. Eventually, the mechanics of this attack allowed for a carrousel of transactions, thereby creating an ever growing snowball. Similar analyses, usually based upon insider threats, can also be made for fraud cases such as the well-documented Société Générale case, but also for the WorldCom and Parmalat cases.

Addressing these problems requires, on the one hand, strong security and compliance guarantees. On the other hand, these guarantees must be substantiated by formal methods ensuring a verifiably secure business process enactment. It should be noted that these concerns are not confined to the financial service sector or to insider threats. For example, the planned unification of Eurpean data protection law into a sole Data Protection Regulation law is likely to change the statuatory duties of the private sector. Under this plan, companies will be legally required to report any breaches of this regulation and may be liable to penalties in the range of 2–5% of their global annual turnover. European industries seem to be ill-prepared to ensure that their information systems and processes will comply with the security requirements of that upcoming regulation, and the threat of substantial fines means that there is an urgent need to create more resilient systems and processes, which calls for more research within the thematic scope of this seminar.

At the interface of security requirements, business needs, and compliance methodologies we can ask many practically relevant research questions, and their answers are bound to have significant impact in academia and industry alike. Relatively little work has been done, however, on adapting or creating new formal methods with which one can check that processes are compliant with rules, preserve demanded privacy constraints, *and* enforce desired security policies at the same time.

One main purpose of the seminar was to present the state of the art in research within the three communities of Security, Verification, and Process-Aware Information Systems to all three communities in an accessible manner and with a view of identifying important research topics at the intersections of these communities. In addition, that exercise was also meant to explore what strategic activities could help in promoting research at the junction of these communities. This agenda was persued through a mix of keynotes, technical presentations, break-out groups under the WorldCafe method, sessions with free-style discussions, and tool demonstrations.

We now highlight some of the key questions and findings can emerged during that week of work – we refer to the online archive of presentation slides, papers, and abstracts for more detailed discussions and findings. Three action items that seemed of particular importance to the participants were:

1. The need for a classification of security properties that are revelant for process-aware information systems, and an understanding of what formal methods might be able to analyse such properties.
2. The need for a set of concrete examples of business processes that are annotated with security considerations or constraints. These might be examples from the real world that have been sufficiently sanatized and anonymized.
3. The need for a review/survey article on the state of the art in formal methods, written for non-experts and ideally for an audience that deals with security, privacy, audit or business processes.

It was also asked what makes formal methods and tools “practical” in this problem space; their was concern about the scalability of these methods, but also about the considerable effort it would take to transfer foundational tools to real application domains – were such somewhat routine but important transfer may not be supported by standing funding models. Concerns were also voiced about the current research in security and privacy, which tends to ignore recent innovations in process composition, such as choreographies.

Another point of considerable interest made concerned the organization of research in this problem space. At the moment, researchers work on problems within their areas and when they begin to collaborate with people from another area this is then more of a bottom-up process where techniques and tools across areas are combined to see what problems one could now solve. It was remarked that it may often be more effective to take a top-down approach in which key problems of the inter-area domain are first formulated and then researchers from the areas get together and try to come up with solutions that draw from their own tool boxes but that may also invent new tools for the problem at hand.

There was also a lot of discussion about what is so distinctive about *process-aware* information systems, and whether these differences to conventional information systems offer perhaps also opportunities. For example, it was discussed whether there is value in validating such systems at a high level of abstraction without considering how such processes get implemented in IT infrastructures and abstraction layers. The participants had mixed views on such merits but it was felt that validation at that level would be easier to realize and that the identification of weaknesses or vulnerabilities at that layer would no doubt be of value.

Another problem mentioned was the need to support legacy systems, and that this need would not go away. Faced with this, it appears that formal validation techniques will have to be able to reason about composed systems in which some parts only have a somewhat well defined interface, but whose internal behavior cannot be guaranteed or predicted to a good degree.

Finally, it was also noted that some of the research problems that suggest themselves to the specialists may not be issues in the field. For example, we may want trusted system composition across organizations but there may not be the need to formally validate such trust since contractual or other legal mechanisms may be in place that incentivize parties to honor that trust, and that give parties a means of seeking damages in case that trust has been violated. On the other hand, such legal mechanisms may not be adequate in the upcoming Internet of Things were 2-party, end-to-end composition will be the exception and not the norm.

2 Table of Contents

3 Overview of Talks

3.1 Deriving RBAC models from process models and logs

Anne Baumgraß (Hasso-Plattner-Institut – Potsdam, DE)

In process-aware information systems (PAIS) permissions need to be tailored, both to allow legitimate users to perform their specific tasks and to avoid fraud and abuse. Role-based access control (RBAC) is a de facto standard to model and specify access control policies. Although RBAC provides a number of advantages for the management of access control policies, the definition of a specific RBAC model is a complex and time-consuming task for security experts. In addition, a constant evolution of business processes as well as its corresponding permissions result in structures that have changed over time and do not necessarily represent a tailored (desired) RBAC configuration of an organization. With the aim to obtain the current RBAC configurations and further support the definition of desired RBAC models,this talk presents an approach that is able to derive candidate RBAC models from process models managed in PAIS as well as log files recorded by PAIS.

3.2 Better enforce than verify! How to ensure compliance of business processes at runtime

Nataliia Bielova (INRIA Rennes – Bretagne Atlantique, FR)

It is well known that compliance is important for business processes. This talk will advocate the use of runtime enforcement techniques to ensure compliance of business processes. Differently from verification, that is concerned with detection of non-compliant executions of a process, runtime enforcement is aimed at correcting business process while it runs to guarantee the desired behaviour. First, we present the case study – a business process of drug dispensation from the Hospital San Raffaele, and an expected behaviour of this process. Second, we show how different runtime enforcement mechanisms can ensure compliance at runtime. Third, we explain and compare the theoretical models of such mechanisms and formal guarantees they provide.

3.3 Security in the Context of Business Processes: Thoughts from a System Vendor's Perspective

Achim D. Brucker (SAP Research – Karlsruhe, DE)

Enterprise systems in general and process aware systems in particular are storing and processing the most critical assets of a company. To protect these assets, such systems need to implement a multitude of security properties. Moreover, such systems need often to comply to various compliance regulations. In this keynote, we present process-level security requirements as well as discuss the gap between the ideal world of process-aware information systems and the real world. We conclude our presentation by discussing several research challenges in the area of verifiable secure process aware information systems.

3.4 On the Parameterized Complexity of the Workflow Satisfiability Problem

Jason Crampton (Royal Holloway University of London, UK)

A workflow specification defines a set of steps and the order in which those steps must be executed. Security requirements may impose constraints on which groups of users are permitted to perform subsets of those steps. A workflow specification is said to be satisfiable if there exists an assignment of users to workflow steps that satisfies all the constraints. An algorithm for determining whether such an assignment exists is important, both as a static analysis tool for workflow specifications, and for the construction of run-time reference monitors for workflow management systems. Finding such an assignment is a hard problem in general, but recent work using the theory of parameterized complexity suggests that efficient algorithms exist under reasonable assumptions about workflow specifications. We improve the complexity bounds for the workflow satisfiability problem. We also generalize and extend the types of constraints that may be defined in a workflow specification and prove that the satisfiability problem remains fixed-parameter tractable for such constraints.

3.5 The Dos and Don’ts of Business Process Compliance

Guido Governatori (NICTA, AU)

The aim of business process compliance is to ensure that the processes of a business satisfy the relevant legal requirements In the first part of this contribution we propose the semantics of legal requirements in terms of the possible ways in which a process can be executed. In particular we provide a classification of the various types of obligations (and related notions) and what they mean in terms of a business process. In the second part we examine how various logics and logical formalisms address the legal requirement, and their suitability to represent legal reasoning with a particular focus on regulatory compliance for business processes.

3.6 Policy-Based Numerical Aggregation of Trust Evidence

Michael Huth (Imperial College, UK)

The decision to trust and so to allow an action may be informed by many, heterogeneous forms of evidence. Often, such evidence is quantitative rather than qualitative, e.g. the age of some software, the location of a device, the amount of a financial transaction, etc. We present a language Peal, for "pluggable evidence aggregation language," in which the aggregation of such evidence can be specified so that aggregated values can be compared with thresholds. These comparisons are conditions that may be plugged into existing policy languages for access control (e.g., XACML) to enrich them with considerations of risk, trust, cost, etc.. We show how such conditions can generate logical formulas that "compile away" any reference to numerical values yet precisely capture said conditions. Then we give a demonstration of a tool we developed, in which we can subject such conditions to important verification tasks, e.g. vacuity checking and sensitivity analysis of thresholds, by analyzing the aforementioned compilations with the SMT solver Z3.

3.7 Formal Methods with Industry?

Fuyuki Ishikawa (National Institute of Informatics – Tokyo, JP)

In accordance with the active discussions during this seminar, this talk presents our activities with the Japanese industry on formal methods (FM). FM has recently attracted strong interests in the Japanese industry. This talk first presents various kinds of guidelines in active FM promotion to "non-FM people". Our educational program Top SE, an educational course for the industry, also has included intensive support for FM. The program has kept increasing attendees (now 40- per year) from the industry for its 1-year educational course. This talk presents experiences in the program, how FM are taught with real(-like) problems and what the top-level engineers tackle after they learn basics of FM (in the process-aware systems for example).

3.8 Policy Auditing over Incomplete Logs: Theory, Implementation and Applications

Limin Jia (Carnegie Mellon University, US)

We present the design, implementation and evaluation of an algorithm that checks audit logs for compliance with privacy and security policies. The algorithm, which we name reduce, addresses two fundamental challenges in compliance checking that arise in practice. First, reduce operates on policies that quantify over data with possibly infinite domains. Second, audit logs are inherently incomplete (they may not contain sufficient information to determine whether a policy is violated or not), reduce proceeds iteratively: in each iteration, it provably checks as much of the policy as possible over the current log and outputs a residual policy that can only be checked when the log is extended with additional information. We implement reduce and use it to check simulated audit logs for compliance with the HIPAA Privacy Rule. Our experimental results demonstrate that the algorithm is fast enough to be used in practice.

3.9 Checking System Compliance by Slicing and Monitoring Logs

Felix Klaedtke (ETH Zürich, CH)

Many kinds of digitally stored data should only be used in restricted ways. The intended usage may be stipulated by government regulations, corporate privacy policies, preferences of the data owners, etc. Such policies cover not only who may access which data, but also how the data may or must not be used after access. An example of such a usage restriction is that "collected data must be deleted after 30 days and not accessed or forwarded to third parties." A promising approach to detect violations with respect to such policies in IT systems utilizes monitoring techniques.

In this talk, I will present a scalable solution for compliance checking based on monitoring the agents' actions, where policies are specified in an expressive temporal logic and the system actions are logged. In particular, our solution utilizes the MapReduce framework to parallelize the process of monitoring the logged actions. I will sketch the theoretical framework underpinning our solution, i.e., the sound and complete reorganization of the logged actions into parts, which we call slices, and that can be analyzed independently of each other, and orthogonal methods for generating such slices, and means for combining these methods. Finally, I will report on a real-world case study, which demonstrates the feasibility and the scalability of our monitoring solution.

3.10 Information Flow Security for Business Processes – just one click away

Andreas Lehmann (Vattenfall Europe Netzservice GmbH, DE)

Introducing Anica as automated non-interference check assistant and Seda to do non-interference checks in business processes in a user friendly way.

3.11 Correct PAIS – A Naive Academic View

Niels Lohmann (University of Rostock, DE)

In my talk, I wanted to give an impression how academia copes with the verification of process-aware information systems. To achieve correctness, we can try to proof that a model satisfies a specification. As an example, I summarized work to verfiy information flow control for business processes. By transforming the non-interference property into a reachability problem, we can use existing algorithms and tools.

In a different school of thought, correctness by construction aims at generating a satisfying model out of the specification. To exemplify this, I sketched how models compliant to legal regulations and role-based access control can be synthesized automatically.

Both examples are supported by tools which scale well on industrial business process models. This encourages the hope of integrating verificiation techniques into practical tools in a spellchecking-style.

3.12 Data-Aware Business Processes: Formalization and Reasoning Support

Marco Montali (Free University of Bozen-Bolzano, IT)

The need for overcoming the traditional dichotomy between data and processes has been extensively advocated both by industry and academia. However, the "data-process engineering divide" still affects the majority of contemporary process-aware information systems. In this talk, we will review our recent research activities, whose common denominator is to demonstrate that making business processes data-aware paves the way towards a better understanding of process-aware information systems. Despite undecidability issues, we will show that it is possible to isolate interesting classes of data-aware processes for which reasoning support is indeed feasible. We will mainly focus on formal verification, synthesis and data quality concerns, giving also a glimpse on monitoring and mining.

3.13 Automatic Analysis and Certification of Policy Safety

Charles Morisset (Newcastle University, UK)

Attribute-based Access Control (ABAC) extends traditional Access Control by considering an access request as a set of pairs attribute name-value, making it particularly useful in the context of open and distributed systems. However, ABAC enables attribute hiding attacks, which allow an attacker to gain some access by withholding information. In this talk, we revisit in the context of the language PTaCL the notion of safety introduced by Tschantz and Krishnamurthi, and we present the tool ATRAP (Automated Term Rewriting for Authorisation Policies), which allows for the automated analysis of safety for PTaCL policies, by producing counter-examples or generating Isabelle proofs of safety.

3.14 Trading Efficiency/Decidability for Expressiveness: Architecture Modeling with Fitzroy

Michael Norrish (NICTA – Canberra, AU)

The Fitzroy system provides a rich language for specifying component-based architectures. In emphasising data types such as lists and records, as well as full-blown arithmetic, we embrace undecidability, hoping that our reasoning support will work sufficiently well in pragmatically interesting cases. Our primary application is now modelling the architectures of systems built on top of the seL4 microkernel. Given that kernel's features, we expect many systems to be particularly concerned with security properties, and we have performed preliminary taint analyses to derive information flow properties of some sample system designs.

3.15 Process Mining: Discovering Process Maps from Data

Anne Rozinat (Fluxicon Process Laboratories, NL)

Most organizations have complex processes that are invisible, thus hard to manage or improve. Each stakeholder sees only part of the process. Manual discovery through workshops, interviews, and review of existing documentation is costly and time-consuming, and rarely reflects actual process complexity. Process mining closes this gap by making the real process visible. Our process mining software Disco leverages existing IT data to generate a complete, accurate picture of the process, with actionable insight. Disco automatically analyzes actual process flows, highlights bottlenecks, shows all variants, and allows animated "replay" of the process flow, all done interactively, driven by process questions.

3.16 Interval-based Process Monitoring for Uncertain Event Streams

Matthias Weidlich (Technion, IL)

Run-time monitoring of process execution is an important feature of process-aware information systems to satisfy security and compliance requirements. In this talk, we focus on two aspects of run-time monitoring, i.e., interval-based event semantics and uncertainty management. We argue that the lifecycle of business activities suggests to consider segmented interval events as the underlying model for reasoning. Further, we present three probabilistic methods for reasoning about constraints defined for segments of interval events for uncertain event streams for which the time of occurrence is unknown for certain events.

3.17 Preserving Demanded Privacy Constraints

Edgar Weippl (Secure Business Austria Research, AT)

The collection, processing, and selling of personal data is an integral part of today's electronic markets. However, the exchange of sensitive information between companies is limited by two major issues. Firstly, regulatory compliance with laws such as SOX requires anonymization of personal data prior to transmission to other parties. Secondly, transmission always implicates some loss of control over the data as further transmission is possible without knowledge of the data owner. In this paper, we introduce a novel approach that aims at solving both concerns in one single step, thus, our k-anonymity-based algorithm is at the same time able to anonymize and fingerprint microdata such as database records. Furthermore, we show that both the anonymization strategy as well as the fingerprint are collusion-resistant, which means that a group of data receivers is not able to subvert neither of the properties by combining their data sets.

3.18 Architecting with Architectural Design Decisions in the Context of Verifiably Secure Process-aware Information Systems

Uwe Zdun (University of Vienna, AT)

In the recent years, the software architecture community has proposed to use architectural design decisions (ADDs) for capturing the design rationale and recording the architectural knowledge (AK). While this approach helps to address key problems in architecting software systems like the gradual vaporization of AK over time, there are a number of important factors that hinder the widespread application of the approach. Examples are the considerable effort required for documenting ADDs in many existing approaches, the lack of integration with architectural design methods, and the limited understanding of how architects decide and understand software designs. The talk will introduce the topic and discuss key open challenges in the area of architecting with ADDs and present results from recent research in this area, especially with regard to and in the context of verifiably secure process-aware information systems.

3.19 Predictive Security Analysis Runtime

Maria Zhdanova (Fraunhofer SIT, DE)

Security Information and Event Management (SIEM) is a key technology to obtain a holistic view of an organization's security state and identify (emerging) security threats for timely and adequate response to security incidents. Today's service infrastructure paradigms such as IaaS and PaaS raise new challenges for SIEM solutions in terms of multi-level/multi-domain security event processing and predictive security monitoring. The EU FP7 IP project MASSIF (Management of Security information and events in Service InFrastructures) aimed to create a next-generation SIEM architecture addresses these challenges in relation to four industrial scenarios: Olympic games, mobile money transfer service, critical infrastructure process control (dam), managed enterprise service infrastructures. Security monitoring in MASSIF SIEM is performed by the Predictive Security Analyzer (PSA) that implements predictive security analysis at runtime in order to identify misuse patterns in event streams. Predictive security analysis at runtime is a novel method for the evaluation of security-related events and their interpretation with respect to the known process control flow and given security properties. Based on real-time events from the process execution environment, a formal process process (APA) and security model (monitor automaton) the PSA allows to predict the close-future process behavior and detect possible violations of security requirements even before a critical situation occurs. One of the misuse cases that demonstrates the applicability of the PSA refers to the insider threat in a hydroelectric power plant.

Participants

- Rafael Accorsi
Universität Freiburg, DE
- Lujo Bauer
Carnegie Mellon University, US
- Anne Baumgraß
Hasso-Plattner-Institut – Potsdam, DE
- Nataliia Bielova
INRIA Rennes – Bretagne Atlantique, FR
- Achim D. Brucker
SAP Research – Karlsruhe, DE
- David Cohen
Royal Holloway University of London, GB
- Jason Crampton
Royal Holloway University of London, GB
- Christopher Dearlove
BAE Systems – Chelmsford, GB
- Guido Governatori
NICTA – St. Lucia, AU
- Christian Günther
Fluxicon Process Lab., NL
- Gregory Z. Gutin
Royal Holloway University of London, GB
- Michael Huth
Imperial College London, GB
- Fuyuki Ishikawa
National Institute of Informatics – Tokyo, JP
- Limin Jia
Carnegie Mellon University, US
- Mark Jones
Royal Holloway University of London, GB
- Günter Karjoth
Hochschule Luzern, CH
- Felix Klaedtke
ETH Zürich, CH
- Agnes Koschmider
KIT – Karlsruhe Institute of Technology, DE
- Jim Huan Pu Kuo
Imperial College London, GB
- Andreas Lehmann
Vattenfall Europe Netzservice GmbH, DE
- Niels Lohmann
Universität Rostock, DE
- Raimundas Matulevicius
University of Tartu, EE
- Marco Montali
Free Univ. of Bozen-Bolzano, IT
- Charles Morisset
Newcastle University, GB
- Alessandro Mosca
Free Univ. of Bozen-Bolzano, IT
- Michael Norrish
NICTA – Canberra, AU
- Andreas Oberweis
KIT – Karlsruhe Institute of Technology, DE
- Alexander Paar
TWT GmbH Science and Innovation, DE
- Stefanie Rinderle-Ma
Universität Wien, AT
- Anne Rozinat
Fluxicon Process Lab., NL
- Thomas Stocker
Universität Freiburg, DE
- Mark Strembeck
Wirtschaftsuniversität Wien, AT
- Meike Ullrich
KIT – Karlsruhe Institute of Technology, DE
- Matthias Weidlich
Technion, IL
- Edgar Weippl
Secure Business Austria Research, AT
- Nicola Zannone
TU Eindhoven, NL
- Uwe Zdun
Universität Wien, AT
- Maria Zhdanova
Fraunhofer SIT – Darmstadt, DE

Report on the Dagstuhl Perspectives Workshop 13342

ICT Strategies for Bridging Biology and Precision Medicine

Edited by

Jonas S. Almeida[1], Andreas Dress[2], Titus Kühne[3], and Laxmi Parida[4]

1 **University of Alabama at Birmingham, Dept. of Pathology, USA**
`jalmeida@uab.edu`

2 **Wiss. Zentrum, infinity[3] GmbH, Gütersloh, Germany**
`andreas.dress@infinity-3.de`

3 **German Heart Institute Berlin, Germany**
`titus.kuehne@dhzb.de`

4 **IBM T.J. Watson Research Center, Yorktown Heights, NY, USA**
`parida@us.ibm.com`

Abstract

The systems paradigm of modern medicine presents both, an opportunity and a challenge, for current Information and Communication Technology (ICT). The **opportunity** is to understand the spatio-temporal organisation and dynamics of the human body as an integrated whole, incorporating the biochemical, physiological, and environmental interactions that sustain life. Yet, to accomplish this, one has to meet the **challenge** of integrating, visualising, interpreting, and utilising an unprecedented amount of *in-silico*, *in-vitro* and *in-vivo* data related to healthcare in a systematic, transparent, comprehensible, and reproducible fashion. This challenge is substantially compounded by the critical need to align technical solutions with the increasingly social dimension of modern ICT and the wide range of stakeholders in modern healthcare systems.

Unquestionably, advancing healthcare-related ICT has the potential of fundamentally revolutionising care-delivery systems, affecting all our lives both, personally and – in view of the enormous costs of healthcare systems in modern societies – also financially.

Accordingly, to ponder the options of ICT for delivering the promise of systems approaches to medicine and medical care, medical researchers, physicians, biologists, mathematicians, computer scientists, and information–systems experts from three continents and from both, industry and academia, met in Dagstuhl Castle for a Dagstuhl Perspectives Workshop on *ICT Strategies for Bridging Biology and Medicine* from August 18 to 23, 2013, to thoroughly **discuss** this multidisciplinary topic and to **derive** and **compile** a comprehensive list of pertinent recommendations – rather than just to deliver a set package of sanitised powerpoint presentations on medical ICT. The recommendations in this manifesto reflect points of convergence that emerged during the intense analyses and discussions taking place at the workshop. They also reflect a particular attention given to the identification of challenges for improving the effectiveness of ICT approaches to Precision and Systems Biomedicine.

Seminar 18.–23. August, 2013 – www.dagstuhl.de/13342

1998 ACM Subject Classification I.2.1 Applications and Expert Systems: Medicine and science, H.3.5 Online Information Services: Web-based services, H.4 Information systems applications, J.3 Life and medical sciences: Medical information systems, etc

Keywords and phrases Systems biomedicine, precision medicine, healthcare-related information systems, biomedical workflow engines, medical cloud, patient participation, ICT literacy

Digital Object Identifier 10.4230/DagRep.3.8.87

1 Executive Summary

Jonas S. Almeida
Andreas Dress
Titus Kühne
Laxmi Parida

Water, water, everywhere, nor any drop to drink. So goes Coleridge's *Rime of the Ancient Mariner*. Until recently, the same went for data: everywhere, but not of much use so far, neither for deriving new medical insights nor for improving medical care.

However, three key developments currently help to overcome this problem: the rapid adoption of electronic medical records [1], the dramatic advances in molecular biology [2], and, just as dramatic, the growing pervasiveness of social computing environments combined with a new attitude towards participatory health management [3, 4, 5]. The result is an exciting medley of initiatives devoted to supporting healthcare related information flow ranging from patient-facing resources such as **PatientsLikeMe** [6] to initiatives such as **MD-Paedigree** [7] (EU's FP7) that provides a physician-centric sort of '**PatientsLikeMine**' analogue addressing treatment choices in paediatrics.

Managing the *creative deconstruction* [8] involved in advancing towards systems medicine requires fundamentally changing the use of ICT in both, healthcare and biomedical research. It requires in particular to take account of the new paradigm of **web-centric computing** which is a basic prerequisite for all these initiatives.

Reflecting these concerns, a Dagstuhl Perspectives Workshop on *ICT Strategies for Bridging Biology and Medicine* was held to discuss a wide range of fundamental and foundational issues. These ranged from architectural considerations to data-access policies including *Open/Linked Data, the Semantic Web, Pervasive Hardware Ecosystems, Medical Clouds, Patient-Participation Frameworks, 'Healthcare* 4.0*', Analytical Tools*, and *Medical Education* Clearly, the required changes can only be achieved by initiatives of a broader scale and scope than what can be accommodated within the existing academic organisations. They need to always involve **all** stakeholders in the healthcare environment. In response to these challenges, the discussions led to the following theses and postulates:

(i) An *open-data policy* for healthcare-related information systems is a fundamental and urgent imperative.
(ii) Following the *business-IT alignment* paradigm [9], healthcare should – on all levels – be supported by secure IT-platforms enabling clinical workflow engines that map healthcare-related processes while integrating pertinent data-analysis, visualisation, and engineering tools.
(iii) Such platforms should also take full advantage of advances provided by *cloud services*, *pervasive computing ecosystems*, and the *semantic web*.
(iv) The *participatory potential* of the Web should be exploited to advance new forms of partnership in the healthcare environment.
(v) The acquisition of *ICT literacy* must become a required part of biomedical education.
(vi) Specifically in Germany, the Bundesnetzagentur should be encouraged to setting up a Working Group *Medizinische Netze* to explore options for a *Medical Cloud* within the German healthcare environment.

References

1 Tracy D. Gunter and Nicolas P. Terry *The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions*. J Med Internet Res 7:1 (2005). DOI:10.2196/jmir.7.1.e3.

2 Susan Desmond-Hellmann et al. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. National Research Council (US), Committee on A Framework for Developing a New Taxonomy of Disease. The National Academies Press, Washington D.C., USA (2011).

3 Wikipedia. http://en.wikipedia.org/wiki/Social_computing

4 Barbara A. Israel et al. *Community-based participatory research: policy recommendations for promoting a partnership approach in health research*. Education for Health, 14:2 (2001):182–197.

5 Melanie Swan. *Emerging Patient-Driven Health Care Models: An Examination of Health Social Networks, Consumer Personalised Medicine and Quantified Self-Tracking*. Int. J. Environ. Res. Public Health 6 (2009):492–525. DOI:10.3390/ijerph6020492.

6 Ben Heywood et al. http://www.patientslikeme.com/.

7 Bruno Dallapiccola et al. http://www.md-paedigree.eu/. The European Commission.

8 Eric J. Topol. *The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care*. Basic Books, New York, NY, USA (2012).

9 Wim van Grembergen and Steven De Haes. *Enterprise Governance of IT: Achieving Strategic Alignment and Value*. Springer, New York Heidelberg Dordrecht London (2009).

2 Table of Contents

3 Overview of the Program

During the workshop, eight specific topics were addressed in regular sessions, complemented with impromptu evening presentations and discussions. For each session, one of the participants served as coordinator and one as rapporteur while every participant was invited to contribute – with some being particularly encouraged to contribute to specific sessions. The rapports for each were collected at http://bit.ly/dagict and formed the basis for this report.

Here is the list of the eight topics augmented by various relevant subtopics and the respective coordinators, contributors, and rapporteurs:

1. **Information and Communication Technology for Bridging Biology and Precision Medicine – Chances and Challenges**
   - From web appliances to web services
   - From home medicine to cloud-based health information systems
   - From laboratory systems to reference genomic atlases
   - Architectures and APIs for user-governed ICT

   Coordinator: Jonas S. Almeida
   Contributors: Hans Lehrach, Wolfgang Maass, ...
   Rapporteur: Mark Braunstein

2. **Big and Heterogenous Data**
   - What can Big Data tell us?
   - Genomic Medicine
   - The integration and modeling of heterogenous multi-omic and imaging data from disease and its implications for diagnosis and therapy
   - Fusing bioimaging data with clinical and molecular information (for enhancing a systems view of disease)

   Coordinator: Joel Saltz
   Contributors: Klaus Maisinger, Stefan Decker, Scott Kahn, ...
   Rapporteur: Alex Pothen

3. **How can ICT help us learning more about disease mechanisms?**
   - Architectures and APIs for user-governed ICT
   - Medical Clouds as platforms for annotation, exchange, and joint interpretation of healthcare data by medical experts (and patients?)
   - Statistics, machine learning etc.
   - Electronic health records

   Coordinator: Bernhard Balkenhol
   Contributors: Eric Neumann, Eric Gordon Prud'hommeaux, ...
   Rapporteur: David Gilbert

4. **Virtualisation, Computation, and AI in Medicine**
   - The status of structured (pathway, model etc) databases
   - The virtual oncology, diabetes, ... patient in medical practice
   - Mechanistic models
   - The vision of ITFoM

- How can we extract information from the scientific literature as well as from 'low-grade information' in the web (text mining, the semantic web in healthcare, search strategies in semantic webs)?
- Virtualisation in drug development

Coordinator: Hans Lehrach
Contributors: Laxmi Parida, Pietro Lio', Joel Saltz, ...
Rapporteur: Andrea Splendiani

5. **Molecular Systems Medicine and ICT I**
   - Assessing emergent properties of chronic diseases and disease mechanisms:
     – The dynamics of disease progression, implications for disease mechanisms
     – Parallel toponome decoding and genome sequencing
     – Cancer and the immune system
   - Family genome and tumor genome sequencing – the use of Next Generation Sequencing and its implications for therapy and disease stratification

   Coordinator: Peter Walden
   Contributors: Walter Schubert, Robert Burk, Markus Löffler, ...
   Rapporteur: Eric Gordon Prud'hommeaux

6. **Molecular Systems Medicine and ICT II**
   - A systems approach to diagnostics (including the use of proteins, mRNAs and miRNAs from blood and from tissue)
   - Biomedical engineering and systems-optimisation strategies in medical care
   - Assays for wellness
   - Precision medicine and guidelines for evidence-based medicine: complementary or incompatible?

   Coordinator: Ina Koch
   Contributors: Anja Hennemuth, Helena F. Deus, Susana Vinga, ...
   Rapporteur: Laxmi Parida

7. **The Stratification of Disease into Discrete Subtypes and its Implications for Science, 'Medical Ontology', Diagnosis, and Therapy**

   Coordinator: Markus Löffler
   Contributors: Mark Braunstein, Eric Prud'hommeaux, Peter Walden, ...
   Rapporteur: Susana Vinga

8. **Does the Potential of ICT in Medical Care Require New Forms and Levels of Medical Training, Medical Data Collection, Storage, and Reproducibility, Clinical Logistics, Clinical Trials, and Patient Participation?**

   Coordinator: Susana Vinga
   Contributors: David Gilbert, Jochen Dreß, ...
   Rapporteur: Ina Koch

4 The Participants' Introductions

The meeting began with the (then present) participants introducing themselves to each other:

Joel Saltz: Works at Emory, is moving to start a Biomedical Informatics Department at Stony Brook's Biotech Dept., and works also as a consultant for pharmaceutical companies.

Susana Vinga: Researcher at the TU Lisbon, works on modelling and control of dynamical systems relating to biological processes, runs projects on pharmacogenomics, computational biology, develops supporting infrastructure for the Biotech Deptartment.

Jonas Almeida: Started out as a plant biologist and then went to chemical engineering, machine learning, biostatistics, and bioinformatics, now applying machine learning to clinical informatics in the clinical context.

Wolfgang Maass: Chair for Information and Service Systems, Faculty of Business Administration, Saarland U, Germany. Developing and employing machine learning, cloud technology, Big-Data management strategies etc., works on embedding information and service systems into real-world systems.

Pietro Lio': Works at Cambridge U (UK) on genetics and complex systems, integrating bioinformatics and medical research, aiming to understand *comorbidities* and to link metabolic pathways with disease.

Klaus Maisinger: Works at Illumina, UK. Studied theoretical physics. As the world is now producing large amounts of sequence data (e.g., the National Health Service is funding full-genome sequencing of 100k patients), he wants to see what to do with these data and how we can you use them to affect clinical practice.

Bernhard Balkenhol: Academic Work on complexity problems in 'classical' information theory, then worked for Telefónica Deutschland, now builds SOAs that allow companies and their employees to deal with each other and their customers in a secure network, wants to investigate how his ICT procedures can be applied in the context of medicine.

Alex Pothen: Works at Purdue U on combinatorial and graph algorithms, sparse matrices, and optimisation models. In bioinformatics, he looked for disease biomarkers from mass spectra and for functional modules in biological networks. Recently, he developed algorithms for registering and analysing cell populations from flow-cytometry data and for dynamically classifying samples in high-dimensional high-throughput data in order to investigate immune-system responses by measuring protein concentrations on cell surfaces. He is also interested in high-performance computing and Big-Data management.

Mark Braunstein: Studied medicine and now works as a Professor of Practice and Associate Director at the Health Systems Institute of Georgia Tech where he is involved in fostering research and community outreach aimed at the wider and deeper adoption of health information technology to improve the quality and efficiency of care delivery. He developed Ambulatory Electronic Medical Records using the Massachusetts General Hospital Utility Multi-Programming System, started and run health IT over the next 30 years. He is interested in healthcare-process modelling, expects that some Health Information Systems are terrible because they don't involve process modelling, and wants to learn how such systems look around the world.

Jonas: Is there a *natural home* for such process-oriented ICT work?

Mark: Interactive computing is very applicable in this context. Performance folks are all talking about clinical applications. We want to use healthcare-specific modelling language to optimise the surgical units in a new surgical floor some hospital was planning to build.

Ina Koch: Works as the Head of the Bioinformatics Department at the Institute of Computer Science at Frankfurt U, started out as a theoretical chemist, yet already her thesis dealt with a topic from bioinformatics, currently interested in systems biology and medicine, cooperating e.g. with physicians on Hodgkins Disease: We need to be able to deal with raw data like images and mass spectrometry data. She also uses – and recommends to use – Petri Nets for systems for which one does not have sufficiently precise quantitative kinetic data.

David Gilbert: Works as the Head of the *School of Information Systems, Computing and Mathematics* at Brunel University in London, UK, and as co-director of its *Centre for Systems and Synthetic Biology.* Studied Computer Science and holds a PhD from Imperial College where he worked on computational logic and concurrent systems. Current topics include:

- In Computational Biology: *Protein structure at fold level.*
- In Systems Biology: *Static models of networks – metabolic pathways, dynamic aspects of signaling pathways (in e.g. cancer models), modelling qualitative and quantitative aspects of multidimensional and multiscale systems using Petri nets, ODEs, and stochastic systems, employing also model-checking procedures.*
- In Synthetic Biology: *Developing a common language and set of tools for synthetic biology research and, more specifically, modelling and programming cell-cell interaction as well as contributing to Malaria research.*

Motivated by the desire to get even stronger involved with biomedical application, he is also working on

- Freidreich's ataxia, integrating patient clinical data with biochemical data (human and animal models),
- precision-medicine modelling of interventions in terms of diet and genetic alterations,
- and on creating patient- and group-specific models of biochemical etc. processes for better predictions.

Stephan Decker: Director of the *Digital Enterprise Research Institute* at the National University of Ireland, Galway, devoted to semantic-web research in Ireland, turning the web from *a web of documents* to *a web of knowledge.* Tim Berners-Lee's insight was that it's not so much about technology, it's about how to get the rest of the planet to accept and use the technology and how to create a global agreement to amplify our own intelligence by taking a 'social-technical' approach. To learn how to create infrastructure that enables teams to share knowledge, biology is an ideal topic because it is complex and fragmented, and requires collaboration between groups.

Eric Neumann: Vice President responsible for 'Knowledge Informatics' at *Foundation Medicine*, a company that is commercialising cancer genome sequencing, matching results with targeted therapies and clinical trials. Originally a neurogeneticist, he was unimpressed with the single genome sequencing in the 1990's since it was over-sold to "cure all diseases", and decided to work at BBN Technologies on biological modeling platforms and synthetic biology for education goals. He saw the need for scientists to communicate more effectively about they findings, but saw the computer tools lagging behind. With the help of linked data, he is expecting research data will become a commodity for others, and that big data will increase in value once semantics has been included.

Peter Walden: Originally a biochemist (with a PhD in T-cell regulation), he works now at the Charité in Berlin on the cooperation and the interference of the human immune system and human disease, doing research on clinical cases in a clinical environment while trying to bridge research and clinical practice, yet respecting that physicians approach problems not like researchers and biologists, and that e.g. much mice knowledge does NOT apply to humans.

Titus Kühne: Works as a physician at the German Heart Institute in Berlin, studies causal relationships connecting systems biology and evidence-based medical knowledge, wants to know what he does not need to know.

Anja Hennemuth Works at the Institute for Medical Image Computing in Bremen on tissue/organ models based on image data for therapy planning, support for clinical studies, and auto-processing of image data, integrating this with management-system studies, and is therefore interested in how to manage different types of data and how to integrate image data in more complex models.

Jochen Dreß: Trained as physician (orthopaedics), moved to systems-biology modelling with Markus Löffler for 2 years, spent 8 years at a Contract Research Organisation supporting clinical trials, then got more involved in IT and software development and moved to the Cologne Center for Clinical Trials working on data-quality assurance and IT governance moving clinical research closer to care, works now at the German Institute of Medical Documentation building a new department providing access to public-health insurance data for researchers.

Helena (Lena) Deus: Started out as a marine biologist studying amino acids in sea urchins, was hired by Jonas to work on a database for methicillin-resistant Staphylococcus aureus bacteria, then went to the MD Anderson Cancer Center where she worked with oncologists. As IT folks weren't paying attention to their needs and their excel-file based IT infrastructure didn't let them exchange biomarker information, this created an opportunity for PhD work. So she built systems to exchange such information for decision support. Before SemWeb (an open-source library created by the semantic-web community) became available, data capture was based on tables. But the clinical space needed more variability. So, she went to the Digital Enterprise Research Institute (DERI) to work with Stephan on 'Linked Data' – a concept which is not just about building systems, but instead about viewing the whole web as a database: imagine asking a question and getting back answers that were as precise as the question you asked. While 'all models are wrong, but some are useful', SemWeb data enable injecting more knowledge to make models more useful. One of Foundation Medicine's goal is to support physicians wanting to help patients who are too weak for standard chemotherapy.

Mark: A classic problem is that the folks developing the IT systems aren't (currently) based on our conversations here.

Lena: Yes, that is the 'lost in translation' problem. Physicians issue requirements that the IT folks don't understand and . . .

Mark: Physicians at the largest children's hospitals still grab data by hand to put them into excel.

Walter Schubert: Works as a physician, currently being the head of the Molecular Pattern Recognition Research (MPRR) group at the Medical Faculty of the Otto-von-Guericke-University, Magdeburg, teaching histology and mathematics-based data-analysis techniques, investigates proteins as building blocks of larger super-molecular complexes, founded 'Molecular Systems Histology' with a two-page paper describing how one can analyse whole protein systems without destroying the tissue, discovered the physical displacement of muscle tissue by (not attacking) T cells, and wants to learn more about Big-Data management. As he always deemed mathematics to be very important for his studies, he has a long-standing cooperating with Andreas.

Jonas: Some say Big Data are a waste of time. Biostatisticians are unwilling to go to the data.

Joel: Lots of areas are labeled 'Big Data' with a hope of funding. I've worked on spatial/temporal axes in GIS and was asked in this context 'Which methods scale the process of acquiring and assembling multi-resolution sensor data?'. My thesis: 'What is needed there, largely intersects with needs for assembling data for comorbidity discovery'.

Lena: The four Vs determining the *value* of Big Data are *Volume, Variety, Velocity*, and *Veracity*, see e.g. [1, 2, 3].

Joel: Those are necessary conditions, but the subclassifications will be important. You'll be able to build systems that apply to many distinct systems.

Walter: Probably invoking an identity crisis.

Zeng Zhenbing: Works in Shanghai, got his PhD in mathematics in Bielefeld, afterwords worked in the Computer Science Institute of the Chinese Academy of Sciences in Chengdu for 10 years, and then at the Institute for Software Engineering, East China Normal University, Shanghai, for another 10. Research topic: writing algorithms for discovering and proving mathematical theorems automatically. Started with bioinformatics in 2005 when Andreas came to Shanghai to establish the *CAS-MPG Partner Institute for Computational Biology* there. 'For the past years, I have learned a lot of biology, but not yet enough to do research – while, amongst my colleagues in Shanghai, I'm the comparative biology expert.'

Eric Prud'hommeaux: Works as 'data plumber' at the MIT, USA.

Andreas Dress: Studied Mathematics, after PhD and Habilitation in Kiel, he went for two formative years to the Princeton IAS, became founding member of the 'Fakultät für Mathematik' at Bielefeld U in 1969, and began to cooperate closely with Manfred Eigen on mathematical problems relating to phylogenetic reconstruction in 1976.

Laxmi Parida: Works at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA in computational biology and leads a Watson group on'Computational Genomics', with a trend towards solving problems through crowd-sourcing, also working on Pattern Discovery, Design and Analysis of Algorithms, Population Genomics, Next Generation Sequencing (NGS) Analysis, and Bioinformatics.

Robert Burk: Physician, works as a neurosurgeon at Johns Hopkins U, doing also research on the neural system and on the human papilloma virus and cervical cancer using genomic and epidemiological information, the human microbiome, the genetics of prostate cancer, and excessive sweating. Wants to learn more computer science and likes to reflect the practice of medicine.

We were later joined by

Scott Kahn: who earned his Ph.D. in Theoretical Organic Chemistry at UC Irvine, performed post-doctoral research at Cambridge University in the UK, and now serves as Vice President and Chief Information Officer of Illumina – San Diego, USA.

Hans Lehrach: who obtained his Ph.D. at the Max Planck Institute (MPI) for Biophysical Chemistry in Göttingen, worked as a postdoctoral fellow at Harvard University, was one of the first who initiated the human genome project, and now works as Director at the MPI for Molecular Genetics in Berlin with a focus on genetics, genomics, and systems biology.

Markus Löffler: who studied physics and medicine and now works as the director of the Institute for Medical Informatics, Statistics und Epidemiology (IMISE), U Leipzig, with

Figure 1 The *word cloud* for the opening session, reflecting the central focus on data, systems, models, and clinical work.

main emphasis on bioinformatics and computational biology, systems biology modelling, and cancer research.

Andrea Splendiani: who works as a computer scientist at the DERI with interests in the semantic web, open data, ontologies, taxonomies, bioinformatics, and systems biology, focusing on the interface between knowledge representation, user interaction, and analytics, mostly in a biomedical domain.

Marc Van Regenmortel: who obtained his Ph.D. in virology from the Medical School at Cape Town U and now works as emeritus Director at the School of Biotechnology, CNRS, Strasbourg U, on molecular recognition and the immune system.

5 ICT for Bridging Biology and Precision Medicine – Chances and Challenges

(The Discussions on Monday Morning)

Coordinator: Jonas S. Almeida
Contributors: Hans Lehrach, Wolfgang Maass, . . .
Rapporteur: Mark Braunstein

(i) Jonas starts moderating the session

Jonas: There is an issue of definitions: we don't know what this workshop is about, what is Big Data? . . .

People have identified data, not knowledge, as the start of the exercise. Apparently, there's a sharp logistical break between the *Resource Description Framework* (RDF) and the *Web Ontology Language* (OWL). E.g., the term 'same As' in OWL has been considered harmful to provenance [4].

Eric N: So, what's data vs. info?

Jonas: The medical side is ready to listen to you: this year has seen the first board exams for informatics relating to the broad area encompassed by data science, i.e. bioinformatics, computational biology, biomedical informatics, biostatistics, information science, quantitative biology, etc.; see also e.g. [5] for the Proceedings of a meeting on Genome-Era Pathology, Precision Diagnostics, and Preemptive Care.

Another question: How do I deal with governance of data to which I am entrusted or not entrusted?

A short demo shows how dragging data into system loads a data-specific interface and how, clicking on a cell, gives you tools to calculate e.g. volume, distance, etc., while `<view-source>` reveals the injected code.

Jonas: Perhaps we need DSL to convert to this code? I've done some analysis. How do I share that data? I've asked as pathologist and statistician. Now some professional dishonesty: I had an algorithm from Georgia Tech to do numerical processing of differential equations. Half-life of a bioinformatics app is 18 months. Can deliver matlab (like other systems) runtime server for free. But the runtime lengths change and folks need to hunt them down on the web. The MatLAB 2006 RT is the biggest download in our directory. Things like Angry Birds have changed the expectations. Now apps don't know about each other.

Two thirds of the data in Electronic Medical Records (EMRs) comes from Pathdata[1], but only about 5% of Pathdata make it into the EMR.

Now, when will we produce patient-facing interfaces allowing patients to log into the US Department of Health & Human Services, to create a family history, and to import data from social apps? Here, helpful advise can be found in [6], an important paper on fusion tables.

Mark: Soon, 5% of patients may have used already Data Share. The HeathVault, a web-based platform from Microsoft to store and maintain health and fitness information, is a good way to do this. The goal of patients integrating their clinical data makes sense, but it's been hard for my grad studs to even create a Google Health account.

Robert: Once you get invalid data, you've corrupted your value.

Mark: That's assuming the EMR data were sound in the first place.

Jonas: Often the person who understands the data about a patient is a relative. Using the governance model of e.g. Microsoft's HealthVault often is a disservice.

Eric N: The notion of a 'best copy' has been hurting us. Data is NOT a flat file though we often think of it that way. Data have multiple dimensions that need to be captured.

David: Patients look at doctors as well.

Jonas: The patient should be the ultimate governor of her/his own data.

Next, a '23andme' TV ad[2] was shown.

Jonas: The ability of authentication systems to delegate enables us to attach governance to data without moving the data. We can't e.g. guarantee patient anonymity.

Robert: People aren't set up to deal with all this information. In the US, it's very hard to get general care because of the changing systems.

---

[1] Cf. http://www.pds-europe.com/products/pathdata-system
[2] from http://www.youtube.com/watch?v=ToloqU6fCjw&feature=youtu.be

Joel: Some patients will want to know lots, some something, and some nothing. But many folks will need interpretation, transmitting granular and noisy data to patient-facing summaries.

Jonas: I agree, but also to physician-facing summaries because physicians need to see the same thing. I saw 'non-rectangular' patient data magically fixed by IMPUTE[3]. The web has the right characteristics for breaking these black boxes.

Walter: Address discrimination against elderly. Elders may live 80km away from where doctors are, they are excluded from healthcare – no internet, no phones.

Mark: I don't buy most patients getting sequenced and managing their data. Most patients (and physicians) don't know what to do with it. We need more input from the patients. In our breast-cancer project, the patients get a 150 USD android with all the apps they need for providing feedback on whether e.g. the pain medications are sufficient. Regarding the elderly, we try to see patients on a regular basis for managing chronic disease, but there is some hope in home devices.

Jonas: Who behind the system is making the judgment and how will this change decisions in disease management? We did a test with 50 people, editing various rules. The outcome was that knowledge bases have to evolve. The requirements will change with time, law, and person. We left it open and folks said 'this is an invitation to chaos'.

(ii) Now, Wolfgang continued moderating the session

Wolfgang: Modes of identification are also at stake. You can give any kind of object an ID – semantic web people would call it an URL. The miniaturisation of hardware allows not only to take pills that are computers, also the price has fallen exponentially. Capturing ids from the real world went from labels to 2d barcode and now to radio-frequency identifiers (RFID). One of the goals: new information projected into an operating room. Other applications: Intelligent bathrooms that interact with you.
Q1. How to design these types of objects?
Q2. What are the technical architectures that enable it?

Eric P: What are the things that a physician would need to see during a surgery?

Wolfgang: Overall patient information from many sources.

Eric P: Such an information ubiquity is probably meant to inject enough context into the environment so that people can work at a higher level. Could be tested in command centers for emergency response or for search & rescue. Inject code into the sensors to make them smarter. This was used in DOLCE[4], but failed. It is now much more implicit in JavaScript than it was in DOLCE. Here, you can create custom objects like self-controlled plants with personalised irrigation. Now, replace 'plant' with 'patient' (though this statement, taken out of context, may not sound so good). Once planted, sensors communicate with edge nodes by representing data using semantic frameworks that process the data using object-oriented functional computing (as offered by JavaScript); then actuators act based on that information. With a million plants, data streams need to be considered. Computation gets pushed into a master node.

[3] A program for estimating ('imputing') unobserved genotypes in SNP association studies.
[4] Descriptive Ontology for Linguistic and Cognitive Engineering, c.f. http:/en.wikipedia.org/wiki/Upper_ontology_(computer_science)#DOLCE_and_DnS.

Mark: In the US, we have 8M patients with chronic disease who account for 50% of healthcare costs. Putting intelligence into the nodes makes sense. There's a company that took technology out of Argon labs and showed that they can build a model of patients to predict which patients are heading towards problems.

Wolfgang: So, we will collect a lot of data. Why to do it with plants: it is difficult to get data from healthcare. And plants allow testing the architecture.

(iii) Now, Hans continues moderating the session and refers to the ITFoM project [7]

Mark: What's going into the model today?

Hans: We're hoping to sequence across a pathology slide. If you put in the oncogenic RAS mutation, you already get the right results. If you put in the genome, your results get better. By definition, we're sequencing the dominant cell. Ideally you'd want to subdivide the tumor into cell types which can react differently.

Jonas: You enter only first-order kinetics?

Hans: In principle, you could enter anything you know about.

Mark: What's the cost?

Hans: 2 times the transcriptome, about 5K.

Mark: For the patient whose metastasis wouldn't be addressed by the drugs, what was the outcome?

Hans: There was another drug that should have addressed that metastasis which came out one month before she died.

6 Big and Heterogeneous Data

(The Discussions on Monday Afternoon)

Coordinator: Joel Saltz
Contributors: Klaus Maisinger, Stefan Decker, Scott Kahn, ...
Rapporteur: Alex Pothen

(i) Joel starts moderating the session and states:

In Data Science dealing with Big Data and in Biomedical Informatics, we need integrative analysis across multiple time scales, modalities, ...
We need clinical pathologists and computer scientists working together for the analysis of images of cancer pathologies.
We have to cope with

- multiple images across multiple modalities,
- exabytes of data from mice,
- the intersection of omics data, radiology, etc.,
- the commoditisation of imaging and sequencing data,
- in vivo high resolution pathology,
- 3D microscopy imaging,
- animal models,
- reconstruction of cellular structures etc.

Yet, in many application areas of multiscale imaging, there is a common core set of tools that could be used.
Extracting features and identifying clusters, three morphological groups have been identified by clustering data from 200 million nuclei provided by *The Cancer Genome Atlas* (TCGA). The big question is how will we cure cancer, how and when to administer stem cells, how to treat other diseases? Using *Random Forests* to construct an ensemble of *decision trees* for identifying the top ten percentile of high risk patients for readmission (in the next 30 days?), a clustering analysis of data regarding Emory's patients and 180 other hospitals from a University health consortium helped to better learn how to predict readmissions and survival, and how some treatment was working at a given moment. The data included clinical phenotypes (heart disease, ...), geographic distribution, and clinical outcomes. Due to what individual medical coders do, the data were also quite noisy.
There were a few terabytes of data including critical care data from sensors.
There is software now for predictive modelling to support healthcare-data analytics and spatiotemporal data management and analyses.
There is also experimental work employing in vivo microscopy at the MD Anderson Cancer Center on the horizon correlating imaging phenotypes with genomic signatures.

Jonas: It should be worthwhile to pool the data even when there are differences among the treatments. We should design a good physical health environment including smart bed towers able to track locations, analyze streaming physiological data, obtain info on smart phones, tablets, etc.
Where do you put genome information? Better put this in the healthcare data analytics rather than in the physical environment. What about other organisations that track people and obtain streams of sensor data that are reliably and flexibly interconnected?

Andreas: However, whatever hospital you go to today, you'd better have a super immune system!

(ii) After coffee break, Klaus Maisinger and Scott Kahn continued

Klaus: The main challenge of experimental data today is there volume. All data seem to be stored in central repositories.
Question: How will this model evolve? Most probably to distributed storage: keeping track of where the data come from and where they are.
What is the form of data? Many, many files, or linked data?

Jonas: Distributed storage using URLs of the data sources is sub-optimal creating a hodgepodge involving the organisations that pay for data.
TCGA data are not easy to use. The Memorial Sloan-Kettering Cancer Center makes subsets of data accessible. In practice, data is stashed where it is stashed.
And, to report on expensive experiments, caGrid uses Version 4.03 of the Globus Toolkit, produced by the Globus Alliance. It was built on presuming that its data federation would manage data security. It sank on its own weight having just too many features.

Robert: According to Wikipedia, the *Kaiser Permanente Center for Health Research* has 8.9 million health plan members, 167,300 employees, 14,600 physicians, 37 medical centers, and 611 medical offices. An important piece of its overall data-warehouse capacity is its Virtual Data Warehouse (VDW). The VDW is a set of standardised data and methods that enable researchers to produce comparable data, e.g. microbiome genetic data, across collaborating sites for proposing and conducting research.

The VDW is ‘virtual’ in the sense that the raw (‘real’) data remain at the local sites, meaning the VDW is not a multi-site physical database at a centralised data-coordinating center. It might also be called a distributed data warehouse.
What we are doing: coming up with a 10–12 page Dagstuhl Manifesto that could be presented to Government funding agencies, a Manifesto about data, medicine, and ICT. We should think about suggestions for using ICT properly beyond current Health Information Systems. For instance, there is no effort yet to standardise data collections.

Scott: There is a benefit to aggregation. We can store raw data. Should we propose a hierarchy of data products similar to a satellite bus[5]? By design, they provide

- a modular structure supporting a wide range of payload and launch vehicle interfaces,
- a redundancy architecture that can be tailored to meet mission needs,
- and their propulsion can be sized (or eliminated) to meet mission objectives.

Note also that there is the *Electronic Medical Records and Genomics Network* (eMERGE), a national consortium organised by the National Human Genome Research Institute with the task of combining clinical data with genomics.
To achieve interoperability, a higher-level abstraction standard at each institution should get us there about 95% of the time. But we have to look more at the interface. To make healthcare providers better, one needs data and interfaces with tools to use them.

Jonas: I am a bit surprised the linked-data folks in the audience are so quiet.

Scott: For those who are intrigued about this business of sequencing formats embedding the metadata within the file itself: note that FASTQ files from the NCBI/EBI Sequence Read Archive often include a description like e.g.

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

In this example taken from http://en.wikipedia.org/wiki/FASTQ_format, there is an NCBI-assigned identifier[6], and the description holds the original identifier from Solexa/Illumina (as described above) plus the read length. The VCF[7] specification can be downloaded from https://github.com/amarcket/vcf_spec/blob/master/VCFspec_4.2.pdf?raw=true. Yet, there is not yet a clear-cut ‘genomic’ definition of prostate cancer etc. versus breast cancer.

Jonas: Are linked-data models used at Illumina? Which architectures are used for raw big data, just collections of files in some VCF format?.

Scott: Illumina uses data for quality-control purposes as well as to search for allelic variations. A clear distinction is that one has raw data, and data that one needs to do analytics about. Also, confirming a result is different from a de novo result.
I am not sure that we deal with Big Data. We have no need to use Hadoop on a 1000 genomes. We only look at a subset of the data. So, tools for smaller data suffice. Genomic data is clean and consistent compared with clinical data.
Also, setting up a ‘Data Federation’ is not ‘One ring to rule them all’. One needs a strategy to ‘federate’ data. Some data may usefully put into data silos so that comparisons can be done across distributed data sets.

---

[5] Cf. http://en.wikipedia.org/wiki/Satellite_bus
[6] NBCI = National Center for Biotechnology Information.
[7] VCF = Variant Call Format.

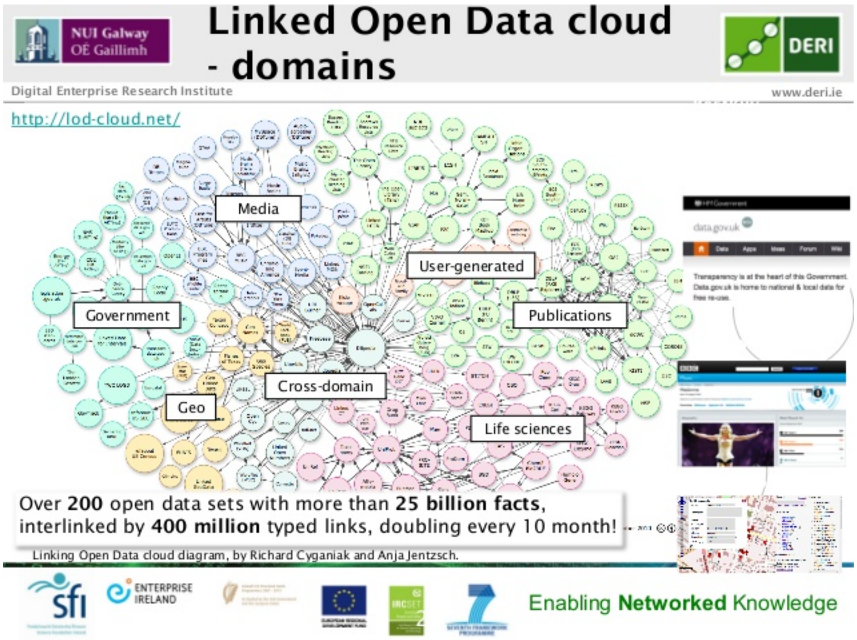


Figure 2 Linked open data cloud domains: a representation of over 200 open data sets.

Jonas: We should start marking data that is least valuable. A good indexing and annotation system could help solve this problem. E.g., we need to distinguish between missing variant transcriptome data, and the absence of variant data.
There is also data heterogeneity as data come from different sequencing methods.
Semantic data is tiny; e.g., the liquid vs solid tumor distinction is based just on where in the body they grow.
Functional genomics is also not yet as applicable as one might hope. E.g., for cutaneous lymphoma, there is no signature mutation present in all tumors. Instead, we have about 25 or so different kinds of such tumors. And all of them are different in different individuals.
Precision medicine and international genome efforts depend on interoperability of data (such as the TCGA data). The report by the National Academy of Sciences entitled *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease* [8] asks to reclassify diseases based on molecular mechanisms that underlie them.
All that is a long way to go.

Robert: Medicine is going through changes, its work is not just disease-based, but also conditions-based. Precision might be unhelpful, and also expensive[8]. And Leroy Hood's P4-Medicine elaboration of precision medicine [9, 10, 11] wants medicine to become predictive, preventive, personalised, and participatory.

(iii) Finally, Stefan Decker presented his views on *knowledge pipelines*

He considers the semantic web rather a *social* than a *technical* initiative.
There are two ingredients:

[8] Four talks on practical efforts in delivering precision medicine can be found at http://www3.ccc.uab.edu/index.php?option=com_content&view=article&id=591&Itemid=300.

Figure 3 The word cloud relating to Stefan Decker's presentation and the ensuing discussions.

1. RDF, the Resource Description Framework, offering a graph-based data representation: URLs are objects, interlink information defines the relationships between objects.
2. Vocabularies (Ontologies): They provide *shared understanding* of a domain, organise knowledge in an *machine-comprehensive* way, and give an exploitable *meaning to the data*.

The linked open data cloud brings together information in various databases as described e.g. in http://lod-cloud.net referring to 200 open data sets (see Figure 2 above).
One goal is to almost 'automatically' transform data into knowledge: There are patients with breast cancer who respond well to the chemotherapy drug *Cisplatin* while others do not. Now comes a third patient: what should the doctor do? The patients' responses depend on their genome and its epigenetic regulation. In fact, there are microRNAs silencing tumor-suppressor genes that can be used to distinguish patients who respond well from those who do not. This could be done in a few minutes instead of weeks.

Stefan also presented a figure of Knowledge Pipelines and the three steps of a Workflow:
Step 1: Load publications into text index to create an overview.
Step 2: Load clinical trial data into RDF database and test queries.
Step 3 : Create a filter on publications for 'neoplasm' and load results into RDF database.

Another useful tool is 'Bio2RDF', a biological database using semantic-web technologies to provide interlinked life-science data supplying a data interface and a query interface designed to help biologists to focus on their discovery process rather than the IT effort.

Finally, Stefan presented a CloudSpace demo.

Eric N: When you make a Solr index[9], you are creating triples?

Stephan: Yes, and then feeding them into the next step of the pipeline (hence 'pipeline').

Jonas: I see how you use SPARQL[10], but will you allow code injection?

Stephan: Not now. Let's work on this together.

Now, Stefan presented a second CloudSpace demo.

---

9 Solr is an open-source enterprise search platform for distributed search and 'index replication', a method for promoting data availability for searching while preventing data loss, cf. http://en.wikipedia.org/wiki/Apache_Solr.
10 SPARQL is a Protocol and RDF Query Language, see e.g. http://en.wikipedia.org/wiki/SPARQL.

Figure 4 The *word cloud* for the session on big and heterogeneous data.

Andreas: Is that a real implementation or a demo?

Stephan: It is based on a script that consumed data from the Linked Clinical Trials data space at http://linkedct.org.

Andreas: You said 'when you take the results', how do you do that?

Stephan: We don't have anything that automagically transforms data, but it sets up, for example, your hadoop structure and stuff like that.

Mark: What's involved in getting the source into RDF?

Stephan: Identify the objects, e.g. – for clinical trials – patients, drugs, ... Once your data are stored in relational form, getting them into RDF format is standard.

Mark: So, if I had thousands of clinical records in SQL[11], 'RDF-fying' them would be serious work?

Eric P: You'd probably leave it in situ and create an R2RML[12] wrapper to provide a SPARQL view.

Wolfgang: How did you solve the vocabulary mapping problem?

Stephan: That came from composing the SPARQL query.

Andreas: The query was in terms of MESH terms in both data sources, so the mapping was 'already done'.

7 How can ICT help us learning more about disease mechanisms?

(The Discussions on Tuesday Morning)

The topics of this session were

- Architectures and APIs for user-governed ICT,

[11] SQL = Structured Query Language, cf. http://en.wikipedia.org/wiki/SQL
[12] R2RML = Relational Database to RDF Mapping Language, cf. http://www.w3.org/TR/r2rml/

- Medical Clouds as platforms for annotation, exchange, and joint interpretation of health-care data by medical experts (and patients?),
- Statistics, machine learning, etc.,
- Electronic Health Records.

Coordinator: Bernhard Balkenhol
Contributors: Eric Neumann, Eric Gordon Prud'hommeaux, ...
Rapporteur: David Gilbert

The discussion started with some general remarks regarding the possible outcome of the conference.

Andreas: Could it make sense to use the Manifesto to push for the installation of a medical cloud in Germany?

Titus: What should our discussion focus on? There are illustrative cases of ICT use regarding oncology as well as neuro- and cardiovascular diseases. And there are various possible directions for ICT to push:
(i) mechanistic modelling (incl. visualisation) to improve the understanding of biomedical insights and their medical implications,
(ii) documenting patient outcome,
(iii) making data & tools accessible to stakeholders.

Next, Joel proposesd three topics on the blackboard:
1. A wider view of health (P4?),
2. Data first vs mechanism/understanding,
3. Data access / sharing.

There ensued a lively discussion on this during which the following statements were made:

A better understanding of disease is an essential component of patient care.

An informatics core common to both, (2) and (3). Yet, how do we make the data available to the different communities?

Grant issues lead to a different focus in different communities.

A better understanding of disease also leverages new science / materials. These have not yet been integrated into practice of medicine. We should focus on this gap.

There is also a gap between disease mechanisms & patients. One needs to make data available. Yet, are we the best consortium to do this?

Informatics, computation, and genomics are driving forces. They will be taken up by medicine in the future. What can we do to improve medicine?

We need to know where we are going. What will medicine look like 10 years in the future.

Titus: Future of medicine: prediction and prevention will be improved.
We need to better understand *pathophysiology* because there are so many diseases that are poorly understood.
Currently, there is interesting work on basic research with a pharmaceutical company at my lab in this direction.
We try out possible therapies on 'model patients'.
There are also fundamental changes in the market place – more smaller pharmaceutical companies. Is there also a change in the economics of the marketplace?

Here are some useful web sites. Regarding

- the current decline in the pharmaceutical industry: see http://www.nature.com/nrd/journal/v11/n3/full/nrd3681.html,
- personalised medicine – its future impact from a pharma perspective: see http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2918032/,
- economics of personalised healthcare and prevention: see http://commonfund.nih.gov/pdf/[13].

Possible conclusion: personalised and precision medicine may be the key to economically develop treatments for diseases for which this is currently not feasible.
Pharmas buy up products from small companies, they have stopped doing basic research.

Eric P: We will have a discussion about the evolution of data standards this evening.

Finally, a video from the EU Flagship Proposal *IT Future of medicine* (ITfoM) was shown.

Tuesday Late Morning discussion:

Mark: We are a bit focussed on e.g. cancer and personalised medicine. But in the developed world, there are problems of ageing etc. which have to be taken into account by policy makers when planning health policies. If we give policy makers the impression that we are just focussed on cancer and personalised medicine, then policy makers will not really appreciate this.
To mention one example: Using modified Markov decision processes, a student made a model of clinical decision-making processes using different modes of care, and compared this with solutions provided by computation. The computer did better than standard models. Modified models have memory and, hence, may make better decisions about treatment.
So, how to design an optimal (real) clinic. We showed that for high-risk patients, the risk can be reduced substantially. We also gave Emory U a different model that could be offered to all 40,000 employees, with a 7% return on investment if the programme would run according to assumptions.

Eric: Are you certain that the assumptions were correct?

Mark: The surgical floor in a hospital can have far fewer beds if configured properly. Yet for stroke treatment, technology has increased costs. So, we need to be careful when suggesting that technology should be used.

David: There are 2 approaches in our discussion:
- data driven [genomics, proteomics etc] which can support basic scientific understanding of disease as well as best-treatment choices, and
- process driven – e.g. healthcare modelling.

Hans: ITfoM is not only just about modelling, but also about making data available. It provides 'continuous' modelling, from the genome of patient to the final outcome and the costs of healthcare.

Mark: So far, the conversation at this meeting is narrower than I would wish.

Others:
- If we make the focus too broad, then we will not make progress.
- Should we sidestep electronic medical records (EMRs)?
- *Enterprise EMRs* will become data repositories.

[13] and http://commonfund.nih.gov/pdf/HealthEconomics_Personalized_Health_Care_summary_REV_9-28-12.pdf for a related summary of health economics.

Figure 5 The word cloud for the Tuesday Morning session.

- We should talk to them, try to liberate healthcare data.
- How to move forward?

Robert: Breakout / work groups?

Andreas: Discuss after lunch.

Bernhard: How should healthcare-related ICT be designed? If you start a project, say, designing a medical cloud, there are several very simple questions to be asked:
- Why do we start the project: Because we would like to change something.
- What would we like to change: The options for hospitals to interact.
- How do we want to do this: By improving communication.
- And so on ...

Data are restricted in use: local, country, international communities. There are resulting restrictions when combining data from different units, establishments, (parts of) organisations, or ....
Functions and/or roles with different security levels need to be conceived, designed, and codified, and rules defining access, activity options, and responsibilities need to be assigned and implemented.
What is the best design of the network topology – should it have a central hub?
How can a new entity join?
Who is responsible to enable access to the network?
Criteria need to be checked, and audits need to be performed regularly.
One option: Construct virtual private networks (VPNs) on application level. Yet who should do this? And how should this be done?

Discussion on different countries' policies regarding the physical location of cloud storage; USA doesn't care, Germany and many others do care.
Most people agree that this is an important issue.

Mark: In the US, there are problems with withheld information exchange. The solution was just to use encrypted emails for data exchange: The US national network for healthcare-data exchange is based on email.

After a coffee break, Roswitha Bardohl (Dagstuhl scientific officer) explains: The Dagstuhl Manifesto should be about 20–25pp long, it must have an Executive Summary and it should cover

- the state of the art,
- visions for the future,
- 6–7 main topics that could form the body of the manifesto.

It should say where research could go in 5–10 years, briefly address open problems. The *Informatik Spektrum* [a german journal] may publish a 2–3 page summary.

Audience: Should it address funding agencies in Germany, but also outside: European Commissions, NSF?
Would Dagstuhl like to collect suggestions regarding people who could / should receive the manifesto?
For the NSF, the executive summary could be 1–2 pages . A 25-pp document would be fine for an NSF working group. Who should be contact people in Germany?
While the *Informatik Spektrum* may publish a 2–3 page summary, should one also try to publish some summary in the US?

After this short explanation, Bernhard continues: The law defines how to deal with personalised data. Someone from security checks users and says whether or not that user is allowed to access certain personalised data. Permission is granted on a case-by-case basis for each data set.
Define what role a person may have at a certain moment and define what the person can do who steps into a role temporarily. Only a person willing to accept a certain role has the right to use some data.
Consider a manufacturing process, e.g. the process of building. There are groups who have rights and perform functions. They have to work out service-level agreements. This implies the need for a central entity for processing registrations.
There is also the problem that different domains in the building can have different languages and may use different words with the same meaning or the same word for different meanings. Can this be solved better by a bottom-up or a top-down service orientation?
We should build systems that are able to learn the meaning of data.
What is the financial consequence if we make changes? Someone may have to pay.
How can I convince two people to change their thinking: why is this or that better? This is the much more difficult problem.
What is specific for a medical hospital? In principle the infrastructure. Changes are data-driven.

Ina: What can we learn from the financial market systems?

Eric P: Discussion on protocols etc.: One should employ RDF.

Hans: George Church [Professor of Genetics at Harvard Medical School and Director of *PersonalGenomes.org*] is getting 1000's of people to put their sequences on-line. RDF documents need encryption, hiding of IP addresses. Yet, if excessive data protection would endangers the life of large numbers of people, this would be bad and would need to be highlighted.

Discussion on data protection, privacy.

Titus: At what time of life do we need access to data? Childhood or later? How often does one need to update the database? Screening – start early. If not done early, one will need to call an institute to do the data collection. Childhood or later? When should we start to trash data in the institute. Data more than 6 months old are not useful to the patient. But keep the genome!

Joel: There is the important history of re-admissions, needs preserving data.

Next, Eric P starts a slide presentation on *The landscape of Healthcare Data* , see http://www.w3.org/2013/Talks/0819-Dagstuhl-egp/, stolen from: http://www.w3.org/2013/Talks/0604-SemTech-egp/.
Standards and initiatives are collected at http://www.w3.org/wiki/HCLS/ClinicalObservationsInteroperability#Standards_Bodies describing

- levels of semantic interoperability,
- policies/initiatives,
- existing standards.

In the bottom-scale 'text soup', we have simple conjunction interpretation; sentiment analysis; NLP (Natural language processing).
At the top, we have complex input and interpretation.
The data size is growing as well as the length of the data vectors (where data are coming from organisations, remote clinical devices, web-based sources, . . .).
Warehouses, in particular clinical-data warehouses show varying coverage and levels of details, use anonymous selection.
i2b2 (Informatics for Integrating Biology and the Bedside) entertains wider networks, yet puts no effort at input coding, lookup tables are built into protocols, garbage in/garbage out.
There are three forms of standards: character level; semantic; semantic web-based.
What to want: Preventable death being prevented, accountability, data for far-sighted decisions, democratised innovation, personalised medicine (cheap enough to save populations of unimportant people), science (e.g. systems biology), P4 medicine [Leroy Hood].
Population wisdom: from bedside to bench and back and beyond.
Population-based conclusions from individual data instances; individual treatment options from population experience; common data appearing in multiple contexts, e.g. syndromic surveillance; outcomes associated with procedures, interventions, and substance administration.

Clinical care:

- US: Affordable Care Act mandates patient portability in the form of Health Level 7 (Consolidated) Clinical Document Architecture (HL7 C-CDA) standards;
- EU: The EU's Innovative Medicines Initiative focuses on clinical data re-use, supporting
  (i) Electronic Health Records for Clinical Research [EHR4CR],
  (ii) the SALUS project on post-market patient safety[14].

Clinical trials:
The European Medicines Agency (EMA) is pushing the publication of clinical-trial data for public health pursueing the goal of understandable clinical data formats.

[14] From the SALUS website: Pre-approval clinical trials cannot guarantee that drugs will not have serious side effects after they are marketed. Post-approval drug safety data studies aim to address this problem; however, their effectiveness is started to be discussed especially after recent examples of drug withdrawals. This is due to the fact that, current post market safety studies largely depend on the submission of spontaneous case reports where underreporting is a major problem. The need for a more proactive approach is apparent, where safety data from multiple sources are actively monitored, linked and analyzed. Effective integration and utilisation of electronic health records (EHR) can help to improve post-market safety activities on a proactive basis.

The US Food and Drug Administration (FDA) pursues the goal of improved use of submitted clinical trial data through 'Therapeutic Areas'.
Pharmaceutical companies have attempted to use submission standards to create study repositories.

Policies – how is it going:
Clinical terms: Too early to tell how SNOMED CT[15] will impact the medical world.
Clinical trials: In 2009, the Clinical Data Interchange Standards Consortium (CDISC) identified a need for intra- and inter-pharma sharing of trial metadata, articulating questions and value sets.
Now, GlaxoSmithKline have developed their own terminology system.
CDISC2RDF, a company developing semantic models for CDISC-based standard and metadata management, now part of the FDA/PhUSE Semantic Technology project, co-operated with AstraZeneca and Roche (Tucson, Arizona) to map SDTM data[16] to RDF while AbbVie, Boehringer Ingelheim, AstraZeneca, BMS, GSK, J&J, Lilly, Pfizer, Roche, and Sanofi formed *TranCelerate* to advance innovation and tackle inefficiencies in R&D, launching three new initiatives and expanding on two existing programmes in its second year of existence.

Clinical informatics: Developing clinical informatics is also considered to be critical to reform, presenting an information model in form of a general-purpose graph structure for medical information on

- activity times and states,
- mood: considered, planned, accomplished,
- factors commonalities in e.g. Prescription vs. Injection,
- acts: observations, substance administrations, procedures, ...
- entities: people, organisations, places,
- roles: patient, nurse, surgeon.

The system tells you where to write, but not how to write defining terminology, data types, and value sets, and uses taxonomies for:

- problem statements,
- tests performed,
- medications administered,
- reactions observed,
- and diagnoses.

Intersecting these, one gets information models, terminology models, and so on that are connected to list complex observations with respective evidence.

The five biggest challenges are outreach & education, architecture, validation, knitting/gluing, and building a physician interface (hardest task!).

Finally, Eric N takes over throwing in a number of unconventional thoughts and catchwords:
There were three important breakthroughs: Kurt Goedel, Ernst Mayr, Eleanor Franklin:

- Kurt Goedel – incompleteness theorems,
- Ernst Mayr – modern synthesis, teleonomic processes,
- Rosalind Franklin (with Watson, Crick, Wilkins) DNA.

[15] SNOMED CT or SNOMED Clinical Terms is a systematically organised computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world.
[16] SDTM = Study Data Tabulation Model, it defines a standard structure for human clinical trial (study) data tabulations.

■ **Figure 6** The word cloud relating to the contributions by the two Erics.

Where are we today? One-directional view of biology: the genomic age – dramatic results. But do they yield the right representation?
Causal formalisms: useful in modelling & predicting biological systems, can also identify flaws & weaknesses in the system . . . .
Yet, living systems are not *control systems*.
The teleonomic principle answers not ‘how’, but ‘why’.
Teleonomic processes owe their goal directedness to the operation of a programme.
Normal evolution (species $\rightarrow$ organism $\rightarrow$ functions).
Basic drivers: energy partitioning, reproduction, mutation, selection (survival).
Cancer emergence (degenerative organismic functions $\rightarrow$ neoplastic functions).
Data $\leftrightarrow$ Models, always separated.
Pathways, systems biology, causal networks.
Meta-mathematics – completeness & consistency.

Basic Diagonal Lemma [key to Goedel’s theorem]. Goedel’s incompleteness theorems: If a system is provably consistent, then it cannot be complete.
What does biological consistency mean?
True statements are unprovable, false statements are disprovable.
The Dilemma. Inconsistency.
ω-consistency: If a theory T is not ω-inconsistent, then it is ω-consistent.
β-consistency: within a biological system S, if a bio-function F defined by S occurs for most inputs x to S, but can be shown not to occur for some inputs z, then it is β-inconsistent.
incompleteness is an immune-related problem. ‘inconsistent’ means that you should not have let it stay.
Immune-related cancer avoidance:
Goedel type 1 errors – biosystem incompleteness [false foreigns],
Goedel type 2 errors – β-inconsistency [false selfs],
goedelian augmentation [GA]. Done by generating new DNA.
Medicine 2020, goedelian loop. Goedel’s incompleteness theorem has possible implications in the limitation of immune and cell regulatory functions

8 Virtualisation, Computation, and AI in Medicine

(The Discussions on Tuesday Afternoon)

The topics of this session were

- The status of structured (pathway, model etc) databases,
- The virtual oncology, diabetes, ... patient in medical practice,
- Mechanistic models,
- The vision of ITFoM,
- How can we extract information from the scientific literature as well as from 'low-grade information' in the web (text mining, the semantic web in healthcare, search strategies in semantic webs)?
- Virtualisation in drug development.

Coordinator: Hans Lehrach
Contributors: Laxmi Parida, Pietro Lio', Joel Saltz, ...
Rapporteur: Andrea Splendiani

To start the discussion, Hans asked:
How much of the way we look at biology is shaped by the way we use to organise research? The publication process likes simple stories (referees don't like the truth: it is too messy and too complicated). Biology is complex: biological systems are built of parts that are selected with no goal or criteria behind it. There is no penalty against complexity.
There are basically two ways to study biology: statistics and modeling.
Modeling has a 'bad reputation' because of some application areas and because of the lack of incentives to validate models in academic research. In 20 years, medicine predicted via analytics will be much better.
Computational approaches to medicine exist in a continuum: from simple statistics to mechanistic modeling based on combining biomolecular insights, genetics and clinical-trial results (e.g. subpopulation performance). As more and more parameters can be estimated, mechanistic approaches will play bigger roles.
One problem is access to parameter values. However, the parameter space can be restricted by observations on model behaviour (feedback). Modeling results should not only be used for prediction. Their deviation from real outcomes should be put back in the modeling process. Models evolve (and their evolution can be to some extent automated).

Andreas: For a good model, systematically comparing the effects of parameter *perturbations* with real outcomes often is particularly useful.

Hans: The medicine of the future will be a self-learning system based on the underlying assumption that all diseases have a mechanistic base.
Of course, not all diseases are amenable to a model-driven approach. Cancer is ideal as you can take samples on which to test a model for intermediate time points. Yet, mechanistic models can also have 'soft parts'.
We have two types of information:

- Open information: what is known about the effects of X on Y (from any source, experimental or inferred). Linked Data is an interesting way to access this knowledge.
- Information that is specific to each patient. Any patient could have an IP number: this would be an interesting way to accumulate information on a patient over her/his

lifespan. There could be several data sources (hospitals, fitbits[17], blood pressure, food, environment, ...). A continuum from medical application to lifestyle.

How do we get all that information together? Not all information is certain.
Artificial intelligence is in a sense artificial. Statistics could have larger weight. The relation between Bayesian statistics and RDF should be explored.

Then, Laxmi started her contribution:
This will be a Big-Data free presentation on *IBM's Watson: The smartest machine on earth*
IBM's *Watson* is good at playing Jeopardy, a game rewarding general knowledge of participants, formulated in complex natural language.
Ultimate test in NLP: won at Jeopardy playing against the best two players in the world. Note that *Watson* had to go through an audition. The audition was not a Turing test (not blind, the presence of the machine could have influenced other participants). After the introduction of machine learning (with baseline NLP + Data Mining), there was a big improvement in performance. It has a massively parallel, probabilistic, evidence-based architecture, required 150 man-years, relies on known facts, and cannot (that is, it was not built to) query the web online. There were three main issues:
Issue 1: Ingest and understand information,
Issue 2: Finding information from a question,
Issue 3: Confidence in answering the question.

Confidence is relevant: wrong answers are penalised.
Watson in summary: Interprets and understands natural language questions, analyzes large volumes of data strored in its memory, generates and evaluates hypothesis and quantifies confidence in its answers, adapts and learns to improve results over time.

To cope with problems in current medicine, IBM is now adapting Watson technologies to healthcare.
Content: Samples of medical texts.
Training: 1300 diagnosis and treatment questions with known answers

History of *Watson*:
- 2006 IBM Research,
- 2011: Jeopardy,
- 2011: healthcare,
- 2012: finances and others.

Watson can also be used to predict the phenotype-genotype mapping, in particular simple-trait mapping.
Evaluation on (part of) the test training set showed a 5–10% of improvement of prediction over the base line which is considered to be a big improvement.
Prediction methods: rrBLUP, a software for genomic prediction, (standard) linear models [untypeable], other standard methods for genomic regression: Lasso, Ridge Regression, Bayes A/B/Cpi, SVM. More exotic methods: Random forests, Markov Nets, Neural Networks, Boosting, ...
MINT: A mutual-information based software for transductive feature selection based on genetic trait prediction, considering interaction between markers (2 markers together may

[17] i.e., wireless-enabled wearable devices for activity tracking, measuring data such as the number of steps walked, quality of sleep, and other personal metrics

■ **Figure 7** The word cloud relating to IBM's Watson.

boost a trait). It led to improvements of 20-25%, may help to find the missing causes for heritability of complex diseases by correlating multi-gene small effects with traits[18].

Mark: Watson is used at the Memorial Sloan-Kettering Cancer Center. It was trained for about 1 year on-site on one disease: adenocarcinoma (all literature + case studies). It is now in clinical use. A doctor sends the entire EMR of a patient to Watson (no genomic data), about 80% is free text. Watson gives a summary (links to papers) and about 60% of the time is confident enough to suggest a treatment. Other facilities in New York are using it as web-server.
The amount of training information specific to a domain is key to confidence (not applicable to all domains in healthcare).

Next, Pietro started his contribution on *From data to models*:

The *Virtual Physiological Human* (VPH) is a methodological and technological framework that, once established, will enable collaborative investigation of the human body as a single complex system.
It is developing a roadmap for the *Digital Patient* and has the potential to fund research projects.
While one needs to integrate all clinical and molecular (omics) data, it is difficult to follow all technologies, and groups will often follow only one type of omics.
The most important information (upstream) is related to DNA conformation (chromosomal territories). Chromosomes change, not only in Leukemia, but also in Diabetes and other pathologies.
It is possible – and may be enlightening – to compute the spatial functional statistics of loci distances in chronic diseases. In this context, tools are important.
Epigenomics reveals heart fatigue. There is a strong correlation between epigenetics and DNA conformation.
Also the gene expression process is linked to epigenetics.

[18] See e.g. Visscher, McEvoy, and Yang: *From Galton to GWAS: quantitative genetics of human height*, Genet Res (Camb). 92 (2010):371-9. doi: 10.1017/S0016672310000571. See also http://www.fimm.fi/en/research/research_groups/group_ripatti/ for a report on studies of Finnish biobank samples that provide a much more precise description of our genomic, transcriptomic, metabolomics, proteomic and other high throughput variation and its relation to complex traits and diseases.

GWAS studies cannot explain more than 10% of the observed variation. A long tail of common variants including more omics data could improve this rate.

Multi-Morbidities: A drug to fight an infection could differ depending on multiple conditions (Metformin targets NKFb. It's pro diabetes, but against Alzheimer). Multi-morbidity makes patient stratification very hard. Omics for diagnostic and prognostic markers needs to be developed.
Classical models are organ-centric. We need to move to process-oriented models.
Processes are mechanistically related. The website http://diseaseome.eu presents maps of diseases related via co-shared pathways.
Important are also multi-modeling methods – including parameter estimation.
There are many modelling approaches: Process Algebra, Compartment-based and Rule-Based Systems, State Charts, Hybrid Systems, Boolean Networks, Petri Nets, Agent- and Lattice-Based Models.
Here is an interesting challenge for multi-scale modeling:
TGFb is anti-cancer under a certain concentration and pro-cancer for higher concentrations (apoptosis inhibitor). Metastasis is first in the bone and produces osteoporosis. Other diseases as well produce osteoporosis.
Our first-order linear model includes a feed-forward loop external to the cell.
Geometry affects modeling. Think of what a few cells that detach integrin can do at the level of the tissue. From that, one can model effects on the bone. One cannot use a single model as there are different time scales.
Model checkers and other formal methods can be used to analyse how a program (model) fits results that can be observed.
What we need is

- multi-omics integration,
- understanding co-morbidity,
- identification of key parameters,
- process modeling,
- causal analysis.

There is resistance toward the adoption of models by clinicians (though examples as insulin pumps prove the contrary). You can imagine that, in a few years, patients could take their own data to train a basic model.
Issue: Where are data for machine learning coming from? There are data that are not in official trials. Results that can be derived are questionable (example: tumor reappearing 4 months after treatment). People debate whether it is appropriate or not to use such data. There is a difference between first-line care and last (least worst) solution. And there are also many other considerations.

Finally, Joel Saltz addressed *Pragmatic Clinical Trials*:

Randomised clinical trials are quite expensive and affected by combinatorial explosion if all factors are to be taken into account (race, age, gender, ...).

Basic concept: If you want to do a clinical trial in reality, you need lots of people, EHRs (not always report forms), even though you may have interest in understanding the impact of a variation of only one factor (A/B comparison), e.g. intervention type (measuring quality of care and other parameters).
Patients populations form a living experimental laboratory.
What sort of recommendation can be given, in particular when you need not only access to data, but access to processes.

Biomedical informatics: In many medical schools, biomedical informatics is done by a group of middle-ranked people with CS skills (should know a bit about machine learning). People doing the technical work are not in the position to change processes (e.g., add a few more questions to be asked).
We need to find mechanisms for making data available for the research side of medical schools (in compliance with privacy) and provide feedback to clinical practice.
PatientsLikeMe: Encompasses 2000 communities about serious diseases. There was a report in the Italian literature that Lithium was helping in a specific disease. They asked doctors for lithium prescription and posted results within 3-4 months on the net (not viable).
Most of privacy laws can be side-stepped if patients volunteer.
Wallgreens, America's big online pharmacy, can find patients matching any clinical trial specification.
Are causal Bayesian models a better statistical model to extract information from 'causal' resources?
One reference for conditional independence testing (à la Judea Pearl who developed Bayesian networks) is http://en.wikipedia.org/wiki/Conditional_independence and on Bayesian networks in general see http://en.wikipedia.org/wiki/Bayesian_networks. And for a reference to the learning healthcare system from the Academy of Medicine, see http://iom.edu/Reports/2012/Best-Care-at-Lower-Cost-The-Path-to-Continuously-Learning-Health-Care-in-America.aspx.

9 Molecular Systems Medicine and ICT I

(The Discussions on Wednesday Morning)

The topics of this session were
- Assessing emergent properties of chronic diseases and disease mechanisms:
  – The dynamics of disease progression and implications for disease mechanisms,
  – Parallel toponome decoding and genome sequencing,
  – Cancer and the immune system.
- Family genome and tumor genome sequencing – the use of Next Generation Sequencing and its implications for therapy and disease stratification.

Coordinator: Peter Walden
Contributors: Walter Schubert, Robert Burk, Markus Löffler, . . .
Rapporteur: Eric Gordon Prud'hommeaux

First, Peter showed two images: Is this an example of a patient who will benefit from intervention?

Question: Patient or model?

Peter: First image was patient, 2nd was a patient-tailored model.

Jonas: What routes do you see for ICT in your work?

Peter: ICT is moving into clinics, but usability is low. So, we may need to breed a new generation of people who can use these tools.
Also, modelling techniques haven't entered the clinic because of usability issues.
The next slide showed heterogeneous cutaneous melanoma with an aggressive immune response attacking the tumor resulting in vitiligo (i.e., whitish skin due to depigmentation), visually-apparent melanoma clones, vascularisation, doomed.

One could not have predicted this just from omics data.
The transformation occurs at the interface between aggressive growth and regression zone. An immune reaction is killing off one clone, thus making space for a more aggressive one.

Jonas: Was having to wait for a physician the mistake? Should we have a home device?

Peter: Absolutely. If your tool causes unnecessary surgery, that's no harm.

On the next slide showing melanoma metastasis, there were far more (tumor-specific) T cells than skin cells. Successful regression is not indicated by more T cells, just by T cell distribution.

On the next slide, the effect of therapeutic vaccination is demonstrated. Deep pigmentation indicates effective immune response.

Lena: There are iPhone apps for sending pics of skin issues (cf. e.g. https://angel.co/goderma-gmbh-3). Are they helpful? What's the balance?

Peter: You can always consult with a physician. We don't know what separates those who will respond to vaccination from those who will not. Anti-tumor responses are, by and large, autoimmune responses.

The next slide shows transplantation-associated cancers. Remarkably, melanoma risk increases 2 to 3 times in transplant subjects. Blood is a poor reflection of what's going on in the tumor. The T cell response is divided into tumor and inflammation response.

Eric N: If there are different groupings, you may see multi-modal correlations after genetic classification.

Peter: Not yet, but tried using many distinct markers. Coinhibitor responses need to be regulated, or you get immune pathologies.

The next slide shows examples of excessive immune response and leads to a discussion of anti-CD28 melanoma trial responses.

Question: What about this claim from Stanford regarding the over-expression of a certain gene in at least 20 kinds of cancer.

Peter: I can only say we know very little about it. We need to integrate omics data with clinical observations. This brings in spatio-temporal resolution, missing from classical omics. Clinical trials are experiments which should feed back into models.

Joel: Lots of folks are working on relationships between omics and pathways. As expected, both are significant. You can get extra precision from omics, but this is currently only good to predict outcome. Characterizing immune response is a 'key 3rd leg'.

Pietro: Predictors of outcome, mortality, and morbidity; what's the difference between omics and classical forms?

Peter: Omics seem to be a tool to get into mechanisms.

Robert: As we move from subjective to objective, we improve. Slides are subjective.

Pietro: If you had a choice between omics vs. patient-participation tools?

Peter: Cancer patients are always willing to participate.

Jonas: Can you push molecular studies to patient devices?

Robert: Sometimes, we can't even do non-invasive glucose monitoring.

Peter: Diagnosis for infectious diseases was always invasive. It moved from spleen biopsy to testing tears.

Markus: Why are we not doing trials on really toxic products? We need a product that makes sense to everyone.

Figure 8 The word cloud relating to molecular systems medicine.

Now, Walter starts his presentation:
We can't draw conclusions on the basis of the data we have. We may need co-mapping to get data, then built a model. There were elegant, logical studies that failed to translate into clinical trials. One needs to investigate mutations and functionality in parallel. Every disease has it's own 'cathedral' (super-molecule).
Next, Walter shows a demo on cross-communication between 100 proteins translated into to sound. The richness [sounds like chaos – Eric P] of the later sample indicates health.
Doing imaging of those 100 proteins also shows reduced functionality in the diseased tissue. Imaging at 1nm resolution revealed 6 layers in separate membranes of which only 3 were known.
An image of a T cell speared by an epithelial neighbor illustrates other remarkable processes. We know by volumetric analysis that killing 300–500g of T cells will kill you.
The whole system matters.

Next, Robert presents his views: Medicine has migrated from historical to evidence-based medicine. We gather a bunch of information going into a cloud. Epidemiologists have pathogenesis knowledge which has developed into action.

Scott: How is that knowledge codified and shared?

Robert: Everyone knows we need to integrate the data and, therefore, need to figure out the impediments: government, insurances and policies direct the data input. We will also have to address the microbiome (totality of bacteria and viruses within our body).

Eric N: We're the transport mechanism for a microbiome.

Robert, presenting a 'heatmap' showing percentages of different reads of body cavities, explains: Advances in microbiome analysis come from NGS. There are two ways to search for it:

- 16S ribosomal RNA in procaryotes, doesn't classify species to phylum, so they talk about Operational Taxonomic Units or 'OTUs'. You amplify a region common to most things and try to guess what's there. Limited to detecting biota already in our database.
- For a gold standard, you can sequence everything.

One can associate ecological measurements of complexity (e.g. Shannon entropy) with disease state. We can handle high variability if it's a strong predictor of disease. Disease in a microbiome is called 'Dysbiosis'.

Scott: Do we know how variable the biome is in healthy people?

Robert: Our microbiome is fairly stable in an individual over time. One is not thinking about simple risk alleles, but instead about interactions.
Regarding cervical cancer, it was thought to be associated with Herpes. It was difficult to prove connection with the human papillomavirus (HPV). With a proper study, the odds ratio is nearly infinite because it's a direct etiological agent. Understanding the molecular mechanism for disease was required for implementing the test.
A study by the *Kaiser Permanente Center for Health Research* involving 1 million people checking for HPV(+/-), Cytology (pap smear), and outcome (CIN3) showed a clear association of cervical cancer with HPV. But it was a great challenge to develop a corresponding effective treatment[19]. The complexity is too much for physician. 20 experts argued about treatment strategy. There is need to use IT to enable physicians without taking away their role as healers, and to query longitudinal data with a better interface to data capture.

Question: I heard that pretty much all males develop prostate cancer, but mostly non-aggressive variants.

Robert: From studies of soldiers, we see lots of stuff that 'looks' like prostate cancer, but we don't know what is going to happen to them.

Mark: The EMR systems were designed with billing in mind, not with what we know we need to know.

Scott: It's not just data capture; two 'epic' systems can't exchange.

Mark: But capture is critical. There are glimmers of hope, but the topic does not get much focus. The systems are ignorant of clinical processes, data modeling, and evidence production.

Finally, Markus takes over:
There are nice models for swarm formation: It suffices to introduce two simple rules to govern behaviour. Gene expression data won't reveal these rules. Models are rules; only informed by data. 'There is no clinical medicine unless in the light of models.'
There is a statistical (non-molecular) model that describes the development of Hodgkin's Lymphoma. Its predictions, validated by a study, established *the* treatment standard that was tried to transfer to non-Hodgkin's Lymphoma.

There are more tested/confirmed models, e.g., the model regarding the growth of epithelial crypts. The old belief was that the large granular lymphocytes are the drivers for the crypt. Our belief: They have no effect, plasticity overcomes. This is supported by a model of exchange between aggressive and dormant lymphona clones.
Personalised medicine started with brain-skull trepanation thousands of years ago. Personalised *molecular* medicine has some good use for splitting into groups, not to get distracted by irrelevant complexity.

Jonas: Suppose someone stratifies your results by similar outcomes, would it be an improvement?

Markus: We need to test if biomarker classifiers are good. There are tests for Burkitt's lymphoma based on *Burkitt-ness classifiers*[20]. However, this detection is not that important because all lymphomas are treated the same way.

---

[19] Cf. *Studies of Cancer in Humans*, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, http://monographs.iarc.fr/EricNG/Monographs/vol90/mono90-7.pdf
[20] Cf. http://www.nejm.org/doi/full/10.1056/NEJMoa055351.

Robert: Regarding such classifications, we may need to have experts curate our inference machines.

Markus: Question: How do we design trials? The *National Cancer Institute* has a protocol for a phase-3 marker test design. In tests that split out subgroups, you need to assure that the performance in derived subgroups is significantly distinct. After implementing a strategy to differentiate folks for whom a particular antibody treatment would apply, we hoped for 10% and were surprised by 20%, highlights the value of this specialisation.

Question: How does the mechanics of the diagnosis work?

Markus: We need molecular testing to discover whether the target for the antibodies exists in the tumor. There is a consistent protocol for producing self-tailored vaccines.
We are currently designing a trial for adaptive HIV therapies (which find the optimal balance between the 30 available single-drug based therapies in order to control the HIV virus and it's many mutations).

Jonas: From the FDA's perspective, your algorithm would be the drug. Do you get a stamp at the end?

Markus: No, they haven't worked out their protocol yet.

Jonas: How do you feel about the data supporting these trials?

Markus: This is not just an observational trial, it's a well-designed prospective cohort study.

Mark: How would you support blinding, i.e., keep docs from knowing they've been augmented by a machine?

Markus: Our plan is to create two teams making recommendations on the same patients. The treatment allocated to the patient is selected randomly from these two recommendations.

Andreas: There is Ingo Althöfer's *Triple Brain* Chess Team[21] involving two chess programs that propose moves and a third person (or program) that selects one of those two. This idea worked quite well.

Markus: It would indeed be good to feed back clinical trial data into a public repository. We've discussed the technical side. We also need to make the algorithms available and need protocols around the biobanks.

Jonas: Would you change the language in a patient consent form?

Markus: Many current consent forms wouldn't enable public sharing. One would need to change language from just 'research' to 'research and ...'.

Robert: The fundamental basis of clinical trials is checking one or two variables. With this, you're introducing many variables. How do you compare them?

Markus: The brute-force design only tests the efficacy of the overall system, not the specifics.

10 Molecular Systems Medicine and ICT II

(The Discussions on Thursday Morning)

The topics of this session were
- A systems approach to diagnostics,
- Biomedical engineering and systems-optimisation strategies in medical care,

[21] Now also integrated in *Shredder*, cf. https://chessprogramming.wikispaces.com/Shredder

- Assays for wellness,
- Precision medicine and guidelines for evidence-based medicine: complementary or incompatible?

Coordinator: Ina Koch
Contributors: Anja Hennemuth, Helena F. Deus, Susana Vinga, ...
Rapporteur: Laxmi Parida

Ina Koch reported on *Some aspects of image analysis of Hodgkin's Lymphoma* and presented some slides explaining her point of view on the subject matter.

Jonas: Your distance is not symmetric. Is that a problem from the biological perspective?

Walter: No. In tumor histology, this is fine. Within abnormal cells, things can be not symmetric.

Alex: How is the distance computed? Euclidian?

Walter: Normal lymphnodes will lead to the same distances.

Ina: Look at which size and shape tumor cells in the neighbourhood exhibit? Still some confusion remains about how the distance is computed.

David: Can one map individuals to these distance maps? Individualised maps?

Ina: Yes. But this is not the main goal. Pathologists still do it by eye.

Andreas: Would you like to eventually replace pathologists?

Ina: NO, we just want to help in complicated cases and learn more about the cell neighborhoods and the distribution of cells.

Anja reports on *From images to models to therapy*:

In contrast to Ina's work, her group uses a top down approach based on *Fractal mathematics*. Can this be used to explain the spread of vasculatures? It seemed that the program could be useful for surgery and, thus, was given to surgeons. So, they became an ICT institute providing surgeons from all over the world with software simulating the vasculature in the liver (virtual liver) to be used for tumor surgery and other therapies. Tissue properties regarding heat distribution were added as well as different biomarkers for stroke etc. to be used for diagnosis and therapy.
As a contribution to system medicine, this went from clinical to complex modeling and personalised and precision medicine[22].

Next, Lena presents slides on *Gene Silencing*.

Eric P: This is just for molecular therapy?? Must have some other uses as well.

Lena: TCGA data grow.

Jonas: Not just 'grows': TCGA data grow exponentially, doubling every 7 months.

Mark: Scale? For 1000 months?

Lena: No, just days, including many preliminary data.

Eric P: Isn't there a large amount of winnowing that happens in the drug industry.

[22] Media attention to this work is documented at http://9to5mac.com/2013/08/21/liver-surgery-now-theres-an-ipad-app-for-that/ and at the *Grand Challenges in image analysis* website http://www.grand-challenge.org/index.php/Main_Page.

Eric N: Pharma companies resist, only use small information at a time, everything has to be presented.

Susana: There is an important reference IEEE paper on *Grand Challenges in Interfacing Engineering with Life Sciences and Medicine* [12]. It emphasises that both, mechanistic and learning models, are important.

Marc: It is indeed a great paper, systems-engineering people are nice – they are problem solvers. It presents the medicine institutes' view as well.

Mark: Point-of-care sensor technologies are a very important and dynamic area.

Robert: Longitudinal studies – a lot of data has been collected, but not used. There is need for new innovative ways.

Walter: Isn't life spatial?

Susana, Joel, Anja: Indeed!

Joel: One needs a combination of radiology, pathology, and omics data to perform *spatial engineering*.

Walter: Does there exist any system that can be used?

Anja: There exist systems for gathering and assimilating such information.

Walter: We need a 3D approach to life. Which system is sufficiently far advanced so that it can be used?

Anja: You can combine our vasculature-systems model with other topics for a modeling approach. The VPH projects aim at this.

Mark, David: Model mashups – different models that come together, cf. http://en.wikipedia.org/wiki/Mashup_%28web_application_hybrid%29!

Mark: Job of systems engineers.

Susana: Largely ignored – ICT is needed for bridging these models.

Mark: THink of a whirlpool engineer, modern washing machines considers chemicals, mechanical etc. aspects. Is there a lesson for biology?

Anja: We are working in this area. Yet, already agreeing on terminology is difficult.

Marc: Sensors are underrated, ON and OFF rates need to be studied. Biological OFF rate can be studied, OFF rate is the time for A and B to stay attached. Then, they dissociate. Should attach long enough and go to the membrane for reaction.

David: There are indeed grand challenges in the IEEE area. Yet, we need to be cautious.

Mark: Progress in the field will require multidisciplinary.

Alex: Flow Cytometry data analysis is a real case of Big-Data robots producing much larger data than can be processed at the moment. They use the same technology as IBM's ink-jet printer from 50 years ago, high dimensional and high throughput, measuring the intensity of the fluorescence (proportional to the number of certain proteins). Instruments are different. There is a problem with continuous data. Compare 'clustering' results across hospitals.

Walter: Different times of the day gives different readings

Alex: Yes, has been taken into account.

Andreas: Regarding the distance between samples: You have several hundreds of samples?

Ina: Do you use different stainings?

Alex: We measure all simultaneously.

Andreas: Do you see any clusters?

David: Or use hierarchical clustering all the time?

Alex: We use hierarchical clustering for community detection (like physicists), the data are collected and aggregated over many cells.

Andreas: Has this still anything to do with medicine?

Alex: Indeed: These are the actual protein measurements.

Joel: This is useful in transplants. A set of drugs that activate/suppress can provide information about compatibilities.

Ina Koch: So, this relates to *complexome* analysis?

Walter: Can the results be compared with normal data? What defines its homeostasis ? Can disbalance be predicted?

Marc: In some, there is age dependency, and in others , there is no age dependency.

Peter: Regarding the composition of the cells, there exist records over a period of time. This could be used for prediction.

Alex: SONY is getting interested in this area. They have announced a new machine that can perform many more observations – since the noise is mixed (gaussian – biological – poisson – electronic), they can measure 50 proteins at a time. The CyTOF Mass Cytometry and Mass Spectrometry platform is monitoring flu spreading geographically. It can also use tissues.

Andreas: Could you pick parts of healing tumors from different regions?

Peter: Cells must be in suspension. Tissues are difficult and tissue cells much larger.

Marc: CD4 is a remarkable receptor in the membrane; HIV attaches itself to this receptor; the number dramatically increases (to 350) and is ready for AIDS – then anti–retro viral drugs interfere. This number is important to decide when an AIDS infection is about to occur. CD4 is a crucial molecule and could be a good target to design a block for attachment to the CD4 target. When antibodies attach, they change the antigen conformation – there is *synergy*.

Alex: In the example above, it should be decreased to 350 cells.

11 The Stratification of Disease into Discrete Subtypes and its Implications for Science, 'Medical Ontology', Diagnosis, and Therapy

(The Discussions on Thursday Afternoon)

Coordinator: Markus Löffler
Contributors: Mark Braunstein, Eric Prud'hommeaux, Peter Walden, ...
Rapporteur: Susana Vinga

Markus: There are not so many prepared talks – anyone wants to contribute?

Mark: Is healthcare at a tipping point? According to Lena's graph, there is a Big-Data revolution in healthcare. Computing stretches from quantitative to qualitative. When you are in a middle of a *disruption*, it is difficult to see it. The first information revolution in medicine was the digitalisation of files.
1864 – Florence Nightinghale
1964 – Lawrence Weed: intelligible hospital records (cf. his paper on 'Medical records, patient care, and medical education)

2001 – crossing the quality chasms – IT must play a central role.

Americans think healthcare is like in TV, US has the best quality worldwide. But: Comparing mortality rates and costs for chronic diseases, it looks not so well.
68%: Chronic diseases costs. Why? Current healthcare delivery system is designed for treating acute illnesses.
For obesity, there are different system of care in the US.
The *Primary Care Physician* – network in the US claims that averaging of multi-chronic disease patients challenges patient records. Life expectancy is higher in Europe.
Federal Approach: Bush appointed some 20 − 30 billion to reimburse cost of EHRs.
Federal Strategy starts with EHR certification for meaningful use (doctors should use them that way). In April 2013, 75% of eligible providers have enrolled, also eligible hospitals are committed to the EHR system.

David: Do they have to put everything in electronic format?

Mark: Not old things. Yet, there are
- opportunities: expand the capacity to generate new knowledge , etc.,
- challenges: no strong incentives, privacy, fragmented EHR,
- perverse incentives: shouldn't be on how many visits or tests.

Shows a map illustrating the geographic distribution of 2012 medicare organisations.
Question: None in Alaska? Answer: Just 3 people there.
Political issues (sausage analogy: if you see how they are made...).
Walgreens (the smartest pharmacy in the states for decades) became the first national ACO (Accountable Care Organisation).
Interoperability, data-standards issues: MD can enter free text and it's exported to *SNOMED*, guarantees accuracy, IBM Watson is already addressed.
Projects/Examples:
- Earlier diagnosis of CHF (cardiac heart failure) by machine learning.
- Artificial intelligence framework combining Markov decision processes and dynamic decision networks to infer the most cost-effective treatment/approach.
- Paediatric chest pain: how to manage and evaluate, infer 'red flags', list of symptoms/features?

MDs echo data, compare younger with older doctors. Information from 2000 patients: what are the most important factors. With logistic regression, the result is simple. Does not work so well for major problems. Only hoisting red flags (based on expert knowledge) proved to be useful after all. Involved are the General Electric disease network, Microsoft Amalga, Optum Labs, Mayo Clinic, ...

David: How is that regulated, e.g. Walgreens' expansion to diagnosing and treating patients for chronic conditions such as asthma, diabetes and high cholesterol.

Mark: Well defined context, training, nurses can do that.

Robert: Be careful: For diabetes: yes, for hypertension: more difficult.
Involved are companies and researchers. The 2*net* device, a wireless hub, measures physiological data from home and sends it to the cloud employing qualcomm chips.
What type of data? E.g. in Georgia Tech, they used radar technology developed by the government (not classified anymore) to collect acustic, respiratory, and movement data without touching the patient.
The algorithm for diabetes was developed by an MD student.

The *vg-bio* company tries to improve healthcare through personalised predictive analytics using a new platform providing the earliest insight into deteriorating patient health. They apply software developed to detect airplane failures now to patients: which of them are the sickest.
Future opportunities:
- 4 levelS: clinical practices (people),
- delivery operations (processes),
- system-structure choices (organisations),
- healthcare ecosystems (society).

There is a Georgia Tech and Emory project on 'Clinic logistic predictions'.
Its *Patient-Centered Medical Home* program compares data collected at home to optimise e.g. the time for the next visit.

Wolfgang: Who designed the process model?

Mark: Students.

Jonas: We need to include process data in the medical record in the future.

Mark: Initially, these records were created only to support billing.

David: These issues should also be discussed with the participants of the security seminar in the other room

Wolfgang: Process exceptions: are they considered, when something goes wrong?

Mark: Yes, they are accounted for.

Joel: There is some experience at Stony Brook on how to characterise processes without black boxes.

Mark: Consider an example: In 60 rooms, there are 60 groups, each is focused on its issue. Then, there is no process as a whole.

Andreas: Bernhard's company's software is addressing *processes*, more specifically, business processes.

David: Focused on US processes. How are things in Europe?

Marc: France: Our card has all medical data and is totally portable.

Robert: All Scandinavian countries and biobanks work that way.

Mark: In the US, the federal system makes this more difficult.

Markus: Are there models/prototypes of processes that could be published?

Mark: Yes.

Markus: Even if there are companies involved? As we have Virtual Physiological Human, would it make sense to have also Virtual Hospital Process?

Jonas, Mark: There is *process mining*. However, data need to be free. This should be a key point for the manifesto.

Eric Prud'hommeaux starts by presenting slides on information models used in Clinical Informatics:
- the Health Level 7 (HL7) *Reference Information Model* (RIM):
- http://www.hl7.org/implement/standards/rim.cfm,
- the Continuity of Care Document (CCD): what institutions need to report: http://en.wikipedia.org/wiki/Continuity_of_Care_Document,
- the Clinical Document Architecture (CDA): http://en.wikipedia.org/wiki/Clinical_Document_Architecture.

Example: Description of a clinical record under XML.
No open access to data: Georgia Tech, Emory, . . . , or other consented research data from patients.
He mentions various initiatives providing software for converting XML to RDF called *XML2RDF* or *xml2ref*, and refers in this context to the *Reference Information Model* and the National Academy of Engineering (NAE) report on *Engineering the Health Care Delivery System* that can be downloaded from https://www.nae.edu/File.aspx?id=7417.

Stefan Decker: Starts a short discussion about the role of abstraction in the life sciences based on a number of observations and a conversation with Marc. Stefan says that, in Marc's presentation, the reduction to very few and often just one causation and the lack of abstractions was mentioned which could help to focus on observable behaviour instead of getting lost in too many factors.
Stefan mentions that CS likes to generate abstraction layers in order to manage complexity. These abstraction layers are not naturally given, but usually artificial constructions aiming to reduce the cognitive load of computer scientists (or programmers) when looking for solutions. He suggests that, maybe, the life sciences could benefit from similar abstraction layers that focus on useful behaviours and functions, and make them available as a concept for e.g. data integration and analysis.

Thanks to the efficient handling of Skype technology by the Dagstuhl offices, we could close the day with a Skype discussion with Lee Hood (Seattle) on Education and Training. Lee begins by explaining

The Systems Medicine initiative is being developed to reach all the way to healthcare, and seeks to validate a new paradigm that includes physicians and other healthcare agents. Taking the bottom-up route is advisable in this context. MDs are usually conservative.

Andreas: We need the medical community to support these news ideas, the question is how (explains the goals of the planned Manifesto).

Lee: 6 to 7 years ago, I contacted top medical schools including Johns Hopkins: There were mixed reactions towards training.

David: We understand the paradigm change of P4 medicine. Yet, writing just a white paper is not enough to change how a medical school teaches.

Peter: Explains plans of the Charité in Germany.

Lee: Who is going to be the leader of the (new) Institute will be critical when creating such a new institute.

David: Initiatives in the UK are directed towards medical information systems.

Lee: That is a very important component, particularly if it includes the various omics. The challenge there is to make sure that all of these data, including patients, falls in the right 'fold'. IT for healthcare is enormously welcome, as the Obama Health Care reform demonstrates. Based on a 3-page document, I am scheduled to talk to a parliamentary group in the UK later this year.

Andreas: Can you us send that 3-page document?

Lee: Major changes in health IT requirements are absolutely needed to catalyze these changes. IT for healthcare is the key component in the engineering of this data-acquisition exercise to identify stratification in the patient population at the molecular level.

Andreas: Can you tell us about medical clouds?

Lee: The first stage of data aggregation is to collect data from individual patients even if health IT may not be ready to handle it right now.

Andreas: We have semantic-web experts here. They think that they can handle this aggregation even though this is beyond current health IT.

Stephan then made a summary of the challenges the semantic web faces in bringing institutions to what is the edge of a new data enterprise.

Eric P: The W3C standards initiatives (cf. http://www.w3.org/TR/) work on *Open Web Platforms* for application development that have an unprecedented potential to enable developers to build rich interactive user interfaces buttressed by vast data stores that will be available on any device. He mentions that patients are part of this process and that standards are in fact advancing in this direction.

Lee: Teaching the CIOs about the advantages of these approaches will trickle to the CEOs and others. Patient advocacy and social networks also offer key mechanisms to educate institutions in moving towards systems biomedicine!
Initiatives such as PatientsLikeMe are key in this regard. This was very clear in the AIDS case where new treatments were driven by the patient associations, not by the medical institutions which, in effect, opposed them.

Eric P: Second-level underwriters play a role here, too, because they make care cheaper. How can they be brought to the table?

Lee: The insurance systems in the EU and the US are very different. In the US, they are going to be squeezed by the healthcare reform. So, their participation in a systems-biomedicine initiative would have to be driven by a clear cost-saving proposition. Maybe starting with small insurance providers is a good way to go.
Let me tell you about our 100M initiative in Luxembourg where this is supposed to be advanced: Bringing people to Luxembourg is not ideal because of locally diffuse healthcare systems and the lack of a Medical School. So, alternatives are being considered. On the other hand, the size of Luxembourg is appealing. So, that is also something that creates unique opportunities. Maybe, the Manifesto should mention these issues and the opportunities to advance systems approaches to healthcare.

David: What is systems medicine?

Lee: It's a systems approach to medicine and to disease. Disease is caused by two types of processes. One is genetic in nature, the other is environmental. Almost all processes leading to chronic diseases display a mixture of these two. Another aspect of systems medicine is that the patients will soon be surrounded, each of them, by *data clouds*.
Generating the analytical tools that can reach the healthcare environment is an absolute necessity. Machine-learning approaches capable of doing a good job at extracting signal from noise is another critical component of systems medicine. The network of networks that are documented by the patient-data cloud goes all the way to the social networks that connect people. When folks are perturbed by diseases, all sorts of subnetworks get involved in both, understanding them and treating (or preventing) them. This is a relevant discussion to have.

Stefan: There is a requirement and need for data integration. How do we generate knowledge? Play with and explore the data. This needs to be developed.

Lee: Our Institute has many threads of pipelines. Challenges will be to convince the society that this will be society data (for grandchildren): If the society generates the tools, it is the patients' obligation to give back that information to generate better medicine in the future.

Andreas: Pietro Lio' here works on co-morbidities

Lee: Disease classification (as proposed by the report of the National Academy of Sciences mentioned above) does not make sense and does not take into account systems-medicine approaches.

Figure 9 The word cloud relating to medical education.

Walter: In the fluorescence-microscopy based *Toponome* approach, we take account of the spatial distribution of dozens of proteins in a single cell (or a small piece of tissue). From what we learned, we became convinced that this will be basic for any systems-medicine approach as well as for disease classification. In one specific case, our approach revealed 6 layers in separate membranes of which only 3 were known.

Lee: Studying molecular networks is revolutionising single-cell analysis.

12 Does the Potential of ICT in Medical Care Require New Forms and Levels of Medical Training, Medical Data Collection, Storage, and Reproducibility, Clinical Logistics, Clinical Trials, and Patient Participation?

(The Discussions on Friday Morning)

Coordinator: Susana Vinga
Contributors: David Gilbert, Jochen Dreß, ...
Rapporteur: Ina Koch

For part of this session, we were joined by the participants of the security seminar that met in Dagstuhl for the same week.

Susana starts the discussion with a presentation of Carol Goble's ISMB 2013 slides[23] entitled *results may vary – reproducibility, open science and all that jazz.*

Robert: Recall the rather harsh Flexner Report from 1910. The Report called on American medical schools to enact higher admission and graduation standards, and to adhere strictly to the protocols of mainstream science in their teaching and research. Many American medical schools fell short of the standards advocated in the Flexner Report and, subsequent to its publication, nearly half of such schools merged or were closed outright.

Eric P: Government driven?

[23] Cf. http://www.slideshare.net/carolegoble/ismb2013-keynotecleangoble

Robert: This was driven by the American Medical Association (AMA): It determined the history of medical education in the US from 1910 to now. They developed a systems-based curriculum with a one-end examination after three years. The curricula in Germany and the US/Canada have four requirements including chemistry, physics, and biochemistry.

Walter: In Germany, a doctorate is not necessary.

Eric P: Students who grew up with computers might want to use it also during their study.

Alex: For many students, the doctorate just opens a second pathway.

Jochen: There should be more science included in the curriculum. This is not a question of technology, but of the way we learn.

Robert: According to Wikipedia[24], evidence-based medicine is the *conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients* or, more specifically *the use of mathematical estimates of the risk of benefit and harm, derived from high-quality research on population samples, to inform clinical decision-making in the diagnosis, investigation or management of individual patients.* So, it needs ICT-literate teachers and students.

Mark: Education of medical students is still mainly descriptive, we have to drive the market to adopt ICT.

Andreas: Yet, one should keep in mind that just understanding and using p-values is not necessarily proof of scientific soundness.

Robert: But, medical students need evidence-based probabilities.

Jonas: Informatics rotations, based on fellowships in informatics, are just starting to teach CS to medical students.

Mark: Formal CS is still more an exception in their education.

Jochen: Special courses are not necessary. Rather, ICT aiming to help patients should be integrated into the normal study.

Eric P: Put CS in different classes. According to a discussion with Jonas, there are three questions for the students.

Robert: We need to transform the CS-based lessons in a more structured way.

David: New forms are indeed needed to transmit the definition of terms: systems biology, systems medicine, healthcare systems, ...

Eric N: So, we need to present sort of a collection of new ways to think: thinking in terms of systems has changed it all.

Mark: So far, a healthcare system has never been engineered, they just came about. But now, we need an engineered healthcare system.

Ina: In chemistry, we have synthetic chemistry and analytic chemistry, and both profit from each other. So, synthetic biology and systems biology should be expected to also profit from each other.

David: Polyclinics and other *points of care* feature specialisations at different levels. In-service training is going to be important.What is going on in Australia, or in developing countries? According to a BBC news-channel report, data science and engineering are producing many

[24] Cf. http://en.wikipedia.org/wiki/Evidence-based_medicine#Assessing_the_teaching_of_evidence-based_medicine

different devices that should be taken account of, and a computer-science orientation will become decisive.

Jonas: How should young people be introduced to ICT? Can you start teaching from Kindergarden to 12?

David: In the UK, education is currently shying away from ICT.

Lena: Kids can learn very early. So, teach ICT to young kids.

Mark and David: We will also have to address computer-human interaction.

Mark: All countries have the same problems, e.g. ageing.

David: So, let us look at the global context, 1st/2nd /3rd world, including China.

Zhenbing: In China, all is very good in the big cities, but not so in the countryside.

Wolfgang: Sometimes, there is more flexible data handling in Africa, and technologies are coming back from the 3rd to the 2rd world.

Jonas: Which healthcare systems will arise in Africa and China?

Mark: Vaccination policies will be crucial.

Jochen: The changing climate brings new diseases.So, we just *need* to interact and share knowledge with 2nd and 3rd world countries.

David: There are many stakeholders, more than one can imagine – however, most are not yet involved.

Eric N: There was a public effort for introducing medical records – once pharmaceutical company got involved. Clinical trials organise stepwise. We have to build mechanism to make sharing easier.

David: Look at drugs: in the sports industry, in personalised medicine, in healthcare, ... Drug provision through the internet affects life style.

Teach formal methods: focus on process stuff (modeling, mining, . . .), workflow platforms, optimisation, event logs.

Wolfgang: Focus on the main process is necessary, you loose flexibility by optimisation.

Mark: Accurate time measurements, sensors, and measurement devices are increasingly becoming important for healthcare. They are already introduced in Scandinavia for personalised drug delivery.

Input from the other group: Early this year, the Newton institute ran a program on *Data Linkage and Anonymisation.*

Robert: How do we figure out risk factors? What is the risk of anonymisation?

David: How can you share data? Already for two data sets, there may be process problems, but also visualisation problems.

Mark: Two anonymised datasets put together reveals all.

Jochen: To avoid risks, we are not allowed to freely mix data in my office. If people need it, they first have to ask for permission.

Jonas: So, one needs to develop a 'culture' on how to deal with anonymity.

Jochen: For gene therapy and personalised medicine, this is indeed quite urgent as reproducibility is a basic principle for every scientific method (cf. Carol Goble's presentation). The Taverna workflow-management system should help turning knowledge into action. Yet, why are there so many retractions of scientific papers.

Data-sharing rules are adhered to by only 50% of journals. The G8 open data charter should provide motivations for science. Sharing results is not a goal in itself. And there have been reasons given why scientists don't want to share their data.

Jonas: Yet, scientific journals are the place where applications are disseminated.

Eric P: There are further problems with runtime environments. One needs to store old versions of e.g. matlab in order to run old code.

Walter: Most software institutions in Germany are funded by the taxpayer. Thus, their products are published as open source and full of bugs. Affects e.g. the biological community who uses this buggy software.

Mark: Greenway Medical Technologies is developing and offering apps that allow users to get access to data. I would like to see a design of an app platform that is ubiquitous – has access to all health data.

Jochen: Look at the Meta-manifesto from Carol Goble (also presented at ISMB 2013 in Berlin).

Robert: Universities want tech transfer via patents, this infringes on academic freedom.

Input from the other group: Algorithms are patentable but not 'copyrightable'; programs have copyright.

Eric P: W3 – specifications are royalty free. Critical algorithms can be used for certain purposes. HL7 is open to avoid the trouble associated with semi-closed knowledge based on software such as the Unified Medical Language System (UMLS), the SNOMED component.

Walter: 'Proprietariness' is a problem, I can't find out what is happening in the software. Is this relevant to our Manifesto? Should we include this as an ethical issue.

Andreas: Maple or Mathematica is proprietary, but we can publish the results obtained by using them.

Susana: What does 'free' mean?

Input from the other group: Brazil moved from Microsoft to the Linux Foundation. This was more expensive, but they got the source code. So, they can be certain that they will always be able to read their documents.

Walter: Software must also be validated.

Input from the other group: Software comes with proprietary formats – and the user is tied in to the software. There is a need for open formats. Open formats are beneficial – and have more business if they are used. The physics community has open access. In our community, editors make fortunes by deciding what will be published, even for PLoS One. Should the community self-publish?

Jonas: Here is Tim Berners-Lee's 5-star system for Linked Open Data:
- 1* = on the web,
- 2* = machine readable data,
- 3* = non-proprietary format,
- 4* = RDF standards,
- 5* = linked RDF.

There is an argument relating to scalability. How many stars should our data have for interoperability. Maximisation of secondary uses is the main motivation for adopting the 5-star system.

Andrea: We need to be clear about our motivation for open data.

Walter: Also ethical issues need to be taken into account.

Jonas: See e.g. the document from the Executive Office of the President: *Open Data Policy – Managing Information as an Asset*[25].

Eric P: Or the Yosemite Manifesto on RDF as a Universal Healthcare Exchange Language[26].

Walter: If patients read their record online without visiting their doctor, this may not be so good, the patients may misunderstand.

Mark: Only 15% of patients understand what they are told by medics. Our Dagstuhl Manifesto could either point to these other manifestos, or create its own set.

Mark: The Office of the National Coordinator for Health Information Technology (ONC) may have some meaningful use here[27].

13 Conclusions

As mentioned already, the discussions led to the following six claims:

(i) An *open-data policy* for healthcare related information systems is a fundamental and urgent imperative.

(ii) Following the *business-IT alignment* paradigm [13], healthcare should – on all levels – be supported by secure IT-platforms enabling clinical workflow engines that map healthcare related processes while integrating pertinent data-analysis, visualisation, and engineering tools.

(iii) Such platforms should also take full advantage of advances provided by *cloud services*, *pervasive computing ecosystems*, and the *semantic web*.

(iv) The *participatory potential* of the Web should be exploited to advance new forms of partnership in the healthcare environment.

(v) The acquisition of *ICT literacy* must become a required part of biomedical education.

(vi) Specifically in Germany, the Bundesnetzagentur should be encouraged to setting up a Working Group *Medizinische Netze* to explore options for a *Medical Cloud* within the German healthcare environment.

References

1 McKinsey Global Institute. *Big data: The next frontier for innovation, competition, and productivity*. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (2011).

2 Ixchel M. Faniel and Ann Zimmerman.*Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse*. International Journal of Digital Curation, 6.1 (2011):58–69.

3 Steve Baunach. *Three Vs of Big Data: Volume, Velocity, Variety*. http://www.datacenterknowledge.com/archives/2012/03/08/three-vs-of-big-data-volume-velocity-variety/ (2012), see also http://dashburst.com/infographic/big-data-volume-variety-velocity/.

[25] Cf. http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

[26] Cf. http://goo.gl/mBUrZ

[27] cf. http://www.healthit.gov/policy-researchers-implementers/meaningful-use-stage-2, also check http://www.ted.com/talks/dave_debronkart_meet_e_patient_dave.html.

Figure 10 The *word cloud* for the combined rapports of all regular sessions and evening presentations, reflecting the central focus on open data, nucleating discussions about medical information systems, and disease modeling.

4 James McCusker and Deborah L. McGuinness. *owl:sameAs Considered Harmful to Provenance*. The Tetherless World Constellation http://www.slideshare.net/jpmccusker/owlsameas-considered-harmful-to-provenance (2010).

5 P. J. Tonellato, J. M. Crawford, M. S. Boguski, J. E. Saffitz. *A national agenda for the future of pathology in personalised medicine: report of the proceedings of a meeting at the Banbury Conference Center on genome-era pathology, precision diagnostics, and preemptive care: a stakeholder summit*. Am J Clin Pathol. 135.5(2011):668–72. doi:10.1309/AJCP9GDNLWB4GACI.

6 Hector Gonzalez et al. *Google Fusion Tables: Web–Centered Data Management and Collaboration*. Presented at SIGMOD'10, June 6–11, 2010, Indianapolis, Indiana, USA.

7 Hans Lehrach et al. http://www.itfom.eu, MPI for Molecular Genetics, Berlin, Germany (2012).

8 National Research Council (US). Committee on A Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. National Academies Press, Washington D.C., USA, 2011.

9 Leroy Hood and Stephen H Friend. *Predictive, personalised, preventive, participatory (P4) cancer medicine*. Nature Reviews Clinical Oncology 8.3 (2011):184–187.

10 Leroy Hood et al. *Revolutionizing medicine in the 21st century through systems approaches* https://www.systemsbiology.org/blog-topics/p4-medicine. Biotechnology Journal 7.8 (2012):992–1001. DOI:10.1002/biot.201100306.

11 https://www.systemsbiology.org/sites/default/files/Hood_P4.pdf

12 HE Bin et al. *Grand challenges in interfacing engineering with life sciences and medicine*. IEEE Trans Biomed Eng. 60 (2013):589–98.

13 Wim van Grembergen and Steven De Haes. *Enterprise Governance of IT: Achieving Strategic Alignment and Value*. Springer, New York Heidelberg Dordrecht London (2009).

Participants

- Jonas S. Almeida
University of Alabama at Birmingham, US
- Bernhard Balkenhol
infinity^3 GmbH – Gütersloh, DE
- Mark Braunstein
Georgia Tech – Atlanta, US
- Robert Burk
Yeshiva Univ. – New York, US
- Stefan Decker
National University of Ireland – Galway, IE
- Helena F. Deus
Foundation Medicine, Inc. – Cambridge, US
- Andreas Dress
Shanghai Institutes for Biological Sciences, CN & infinity^3, DE
- Jochen Dreß
DIMDI – Köln, DE
- David Gilbert
Brunel University, GB
- Anja Hennemuth
Fraunhofer MEVIS – Bremen, DE
- Scott Kahn
Illumina – San Diego, US
- Ina Koch
Goethe-Universität Frankfurt am Main, DE
- Titus Kühne
Deutsches Herzzentrum – Berlin, DE
- Hans Lehrach
MPI für Molekulare Genetik – Berlin, DE
- Pietro Lio'
University of Cambridge, GB
- Markus Löffler
Universität Leipzig, DE
- Wolfgang Maaß
Universität des Saarlandes, DE
- Klaus Maisinger
Illumina – United Kingdom, GB
- Eric Neumann
Foundation Medicine, Inc. – Cambridge, US
- Laxmi Parida
IBM TJ Watson Res. Center – Yorktown Heights, US
- Alex Pothen
Purdue University, US
- Eric Prud'hommeaux
MIT, US
- Joel Saltz
Emory University, US
- Walter Schubert
Universität Magdeburg, DE
- Andrea Splendiani
DERI – Galway, IE
- Marc Van Regenmortel
IREBS – Illkirch, FR
- Susana Vinga
Technical Univ. – Lisboa, PT
- Peter Walden
Charité – Berlin, DE
- Zhenbing Zeng
East China Normal University – Shanghai, CN

Report from Dagstuhl Seminar 13351

Coding Theory

Edited by

Hans-Andrea Loeliger[1], Emina Soljanin[2], and Judy Walker[3]

1 **ETH Zürich, CH,** loeliger@isi.ee.ethz.ch
2 **Bell Labs, Alcatel-Lucent, Murray Hill, US,** emina@research.bell-labs.com
3 **University of Nebraska, Lincoln, US,** judy.walker@unl.edu

Abstract

Coding theory has become an essential ingredient of contemporary information technology, and it remains a fascinating area of research. The seminar brought together 45 high-caliber researchers with backgrounds and interests in various different parts of coding theory. The new area of codes for cloud applications received much attention, but other key areas such as network codes, codes on graphs, algebraic coding, and polar codes, were also well represented and generated lively discussions.

Seminar 25.–30. August, 2013 – www.dagstuhl.de/13351
1998 ACM Subject Classification E.4 Coding and Information Theory, I.1.2 Algorithms
Keywords and phrases Coding theory, codes on graphs, polar codes, network coding, index coding, data distribution, cloud storage.
Digital Object Identifier 10.4230/DagRep.3.8.136

1 Executive Summary

Hans-Andrea Loeliger
Emina Soljanin
Judy Walker

While coding theory has evolved into an essential ingredient of contemporary information technology, it remains a fascinating area of research where many fundamental ideas of information theory and mathematics meet. Indeed, the diversity and profundity of recent new ideas in, and new applications of, coding theory is impressive. The following themes were of primary interest at the seminar:

Codes on graphs include turbo codes, low-density parity check codes, and a variety of similar codes. Due to the recent new idea of "spatial coupling", such codes can now be designed to achieve the Shannon capacity of most communication channels with practical encoders and decoders. Such codes are a perfect nurturing ground for cross-fertilization of ideas between computer science, electrical engineering, and mathematics. The mathematical tools in this area include ideas from graph theory, probability, algebra, discrete mathematics, and statistical physics.

Algebraic coding theory continues to be of supreme theoretical and practical interest. Prime examples of this area are Reed-Solomon codes, codes from algebraic geometry, and codes obtained from algebraically constructed graphs. Recent advances in the field include, in particular, list-decoding algorithms for various classes of algebraic codes. Emerging relationships between this area and codes on graphs appear to be promising for future research.

Polar codes (discovered by Arikan in 2008) are a breakthrough of utmost significance. Such codes are provably capacity-achieving on very many channels with very low-complexity (and very practical) encoders and decoders. These codes rely on a new large-system limit that combines information theory and coding theory more smoothly than any prior coding technique. The investigation of such codes, including their combination with other coding techniques (such as codes on graphs and algebraic codes), is an exciting new area of research.

Network coding aims at improving data transmission (throughput, reliability, latency, etc.) in networks. This area is still quite young, but it has begun to influence the design of methods and protocols of content delivery in the internet. There is a diverse set of network coding problem formulations, and network coding can be (and has been) studied within a number of different theoretical frameworks, such as algebraic, combinatorial, information theoretic, and linear programming frameworks.

Codes for cloud applications are about distributed storage of large amonts of data. Diverse requirements on reliability, access latency, updatability, and repairability pose entirely new challenges for coding theory.

In addition, there were also two talks on topics in coding theory inspired by biology.

The seminar brought together 45 high-caliber researchers with backgrounds and interests in these different areas. The seminar was held in the usual Dagstuhl style, with a rather light program of formal presentations and much room for informal interaction. It was interesting and stimulating to hear of developments outside one's own speciality, and (to the best of our knowledge) all attendants greatly enjoyed the seminar.

2 Table of Contents

3 Overview of Talks

3.1 Gabidulin Codes in Characteristic Zero

Daniel Augot (Ecole Polytechnique – Palaiseau & INRIA, FR)

Joint work of Augot, Daniel; Loidreau, Pierre; Robert, Gwezheneg

We transpose the theory of rank metric and Gabidulin codes to the case of fields of characteristic zero. The Frobenius automorphism is then replaced by any element of the Galois group. We derive some conditions on the automorphism to be able to easily transpose the results obtained by Gabidulin as well and a classical polynomial-time decoding algorithm. We also provide various definitions for the rank-metric.

3.2 Some Problems of Coding Theory Motivated by Coding for Memories

Alexander Barg (University of Maryland – College Park, US)

Joint work of Barg, Alexander; Mazumdar, Arya; Kashyap, Navin; Zemor, Gilles

We consider several coding problems motivated by writing onto memories. The first part of the talk is devoted to coding in the space of permutations with the Kendall tau metric. We discuss distance-preserving embeddings of the Kendall space and their relation to bounds on the size of optimal codes. A tight asymptotic bound on codes is derived. We also present a construction of codes in permutations that asymptotically meet this bound. The construction relies on codes in the conventional Hamming space and can be decoded based on their decoding algorithms.

In the second part we discuss a noise model motivating by writing on the granular magnetic medium. We discuss several bounds on codes as well as on the capacity of the probabilistic model of the "grains" channel. Finally, we introduce a problem in binary coding that deals with errors that do not affect adjacent bits (this is a simplified version of the previous problem, but it is nontrivial in its own right). We point out that correcting nonadjacent errors is combinatorially equivalent to correcting conventional errors, but at the same time that there are several scenarios in which this restriction helps to construct better codes that in the standard case of Hamming errors.

3.3 Efficient Projection onto the Parity Polytope and its Application to LP Decoding

Stark C. Draper (University of Toronto, CA)

Joint work of Draper, Stark C.; Barman, Siddarth; Liu, Xishuo; Recht, Benjamin

When binary linear error-correcting codes are used over symmetric channels, a relaxed version of the maximum likelihood decoding problem can be stated as a linear program (LP). This LP decoder can be used to decode at bit-error-rates comparable to state-of-the-art belief propagation (BP) decoders, but with significantly stronger theoretical guarantees. However, LP decoding when implemented with standard LP solvers does not easily scale to the block lengths of modern error correcting codes.

In this talk we draw on decomposition methods from optimization theory, specifically the Alternating Direction Method of Multipliers (ADMM), to develop efficient distributed algorithms for LP decoding. The key enabling technical result is a nearly linear time algorithm for two-norm projection onto the "parity polytope". The parity polytope is formed by taking the convex hull of all codewords of the single parity-check code. Efficient solution of this projection allows us to use LP decoding, with all its theoretical guarantees, to decode large-scale error correcting codes efficiently. We discuss performance results for a number of long LDPC codes, presenting results on error rates and computational complexity. In comparison to BP decoding we observe that the waterfall of the LP decoder initiates at a higher SNR. In conclusion we present a small modification of the LP decoder wherein by slightly penalizing the linear objective of the LP we close the SNR gap between BP and LP.

3.4 Binary Multiplicative Codes

Iwan M. Duursma (University of Illinois – Urbana Champaign, US)

We describe how binary linear codes with an extra multiplicative structure (to be used for algebraic decoding, secret reconstruction, secure or fast multiplication, etc) are either a Reed-Muller code or the concatenation of an additive subcode over a larger field with a Reed-Muller inner code.

3.5 Codes for Secure Distributed Data Storage

Salim El Rouayheb (Illinois Institute of Technology, US)

Distributed storage systems are now a growing paradigm for providing online storage of data and making it accessible anywhere and anytime. In this talk, I will address the problem of achieving information theoretic security of data in these systems to protect it against eavesdropping and malicious attacks. These systems are dynamic due to storage nodes frequently leaving or joining the system. This creates the challenge of safeguarding the system from an adversary which may come at different time instances to observe and maliciously corrupt the stored data. I will give bounds on the secure capacity, i.e., the maximum amount of information that can be stored safely on the system, and describe secure code constructions that can achieve these bounds in certain cases. An important part of the talk will be dedicated to discussing the numerous problems that remain open in this area.

3.6 Semantic Value of Information: Coding and Decoding Schemes Tailor Made for Image Transmission

Marcelo Firer (State University of Campinas – Brazil, BR)

Joint work of Firer, Marcelo; Panek, Luciano; Ramos Rifo, Laura L.; Pinheiro, Jerry A.

We explore possibilities for coding and decoding considering semantic value for errors, in particular schemes that are tailor made for image transmission. To do so, we introduce a loss function that expresses the overall performance of a coding scheme for discrete channels and exchange the usual goal of minimizing the error probability to that of minimizing the expected loss. In this environment we explore the possibilities of using poset-decoders to make a message-wise unequal error protection (UEP), where using a lexicographic order for encoding. We give explicit examples, done for scale-of-gray images, including visual simulations for the BSMC, exploring the encoding and the decoding possibilities.

3.7 On the MacWilliams Extension Theorem for Poset Codes

Heide Gluesing-Luerssen (University of Kentucky, US)

A weight function on R^n, where R is a finite ring, is said to satisfy the MacWilliams extension property if every weight isometry between codes in R^n extends to an isometry on R^n. After discussing various classes of poset weights we state the main result. It says that a poset weight satisfies the extension property if and only if the poset is hierarchical.

3.8 Partial Spread Codes

Elisa Gorla (Université de Neuchâtel, CH)

Joint work of Gorla, Elisa; Ravagnani, A.

As in the approach by Koetter and Kschischang, we study subspace codes as families of k-dimensional linear spaces over a finite field. Following an idea in finite projective geometry, we introduce a class of constant dimension codes which we call partial spread codes. Partial spread codes naturally generalize the known family of spread codes. We provide an easy description of such codes, discuss their maximality, and explain how to decode them efficiently.

3.9 A Characterization of All Invariant Weight Functions on a Principal Ideal Ring With the Property That All Code Isometries Allow For Monomial Extension

Marcus Greferath (University College Dublin, IE)

Joint work of Greferath, Marcus; Honold, Thomas; Mc Fadden, Cathy; Wood, Jay A.; Zumbrägel, Jens

It has been apparent since the end of the foregoing century that finite Frobenius rings (and modules) are the adequate alphabets for ring-linear algebraic coding theory. For these, it was proven early that Hamming isometries and homogeneous isometries allow for MacWilliams' Extension Theorem, which means that linear code isometries that preserve the Hamming and/or homogeneous weight can be monomially extended to the ambient spaces of the codes in question. The talk at hand starts with a question of similar importance: For which subclass of the class of all finite Frobenius rings can we characterize all weight functions that allow for monomial extension of code isometries. Recent years' work has revealed that this question is anything but trivial, however it turned out that we can answer it at least for the class of all finite principal ideal rings.

The talk is dedicated to the memory of Werner Heise, who discovered the homogeneous weight and thereby vastly influenced foundational work in ring-linear coding theory. He died in February 2013.

3.10 Polar Codes: Finite-Length Scaling and Universality

Hamed S. Hassani (EPFL – Lausanne, CH)

Joint work of Hassani, Hamed S.; Urbanke, Rüdiger

Polar codes achieve the capacity of a wide array of channels under successive decoding. Since the invention of polar codes by Arikan a large body of work has been done to investigate the pros and cons of polar codes in different scenarios. We consider two features of these codes that are central from the practical perspective: finite-length scaling and universality. In this talk, after a brief description of polar codes, we will explain each of these features together with some of the related recent results and open questions.

3.11 On the Second Largest Eigenvalue of an n-regular Graph

Tom Høholdt (Technical University of Denmark, DK)

Joint work of Høholdt, Tom; Justesen, Jørn

We give new lower bounds on the second largest eigenvalue of an n-regular connected bipartite graph. This eigenvalue is important for the minimum distance of the graph codes constructed from the graph as well as for the expansion properties of the graph. The proofs involves the quotient matrix and the eigenvalue interlacing theorem. By similar methods we obtain new bounds on the minimum distance of graph codes, in particular in the cases where the previous bounds are useless.

3.12 Graph Codes on Projective and Euclidean Planes

Jørn Justesen (Technical University of Denmark, DK)

Joint work of Justesen, Jørn; Høholdt, Tom

We study codes constructed from Reed-Solomon codes and bipartite graphs coming from projective and Euclidean planes. The code symbols are associated with the edges and the symbols connected to a given vertex are restricted to be codewords in the component Reed-Solomon code. We give exact formulas for the rates and minimum distances of the codes and discuss systematic encoding.

3.13 On the Interior Points of the Storage-Bandwidth Tradeoff

P. Vijay Kumar (IISc – Bangalore, IN)

It was shown by Dimakis et al. that in an erasure code designed for distributed storage, there is a tradeoff between amount of data storage and bandwidth for repair of a failed node. While this tradeoff is known to be achievable under functional repair, the same does not hold for exact repair. A recent result by Tian established that there exist code parameters under which exact repair is not achievable even in the limit of large block lengths. In this talk, we will review these developments. We will also provide a normalized tradeoff that explains why code constructions for the interior points are of interest and identify a construction that achieves a single interior point. It also improves upon space-sharing in the interior region.

3.14 Some Recent Topics in Coding for Secrecy

Muriel Medard (MIT, US)

In the first part of this talk, we present two examples of recent applications of algebraic codes to secrecy. These codes allow us, under the commonly held assumption of source uniformity, to establish strong information-theoretic guarantees. When encryption is limited to coding coefficients of random linear codes, we show information-theoretic secrecy results for such a construction. In particular, we show that the encoded payload and the encrypted coefficients do not yield information about each other. In our second example, we interpret keys as representing the size of the list over which an adversary would need to generate guesses in order to recover the plaintext, leading to a natural connection between list decoding and secrecy. Under such a model, we show MDS codes can be constructed so that lists satisfy certain secrecy criteria, which we define to generalize common perfect secrecy and weak secrecy notions. In the final part of the talk, we revisit the source uniformity assumption that subtends our analysis, as well as much of information-theoretic treatment of secrecy. In particular, we show that, the common treatment of the elements of the typical set as being essentially uniformly distributed fails when one consider guesswork over the typical set. Such results encourage us to revisit uniformity assumptions in secrecy.

3.15 Linear Codes From Oval Polynomials

Sihem Mesnager (University of Paris VIII, FR)

The main topics and interconnections arising in this talk are symmetric cryptography (S-boxes), coding theory (linear codes) and finite projective geometry (hyperovals). Bent vectorial functions are maximally nonlinear multi-output Boolean functions. Such functions contribute to an optimal resistance to both linear and differential attacks of those symmetric cryptosystems in which they are involved as substitution boxes (S-boxes). In this talk, we firstly show that the o-polynomials from finite projective geometry give rise to several new classes of optimal vectorial bent functions whose components belong to a certain Reed-Muller code. Secondly, we present a general construction of classes of linear codes from o-polynomials and study their weight distribution proving that all of them are minimal weight codes. The second contribution shows that some hyperovals of the projective plane from finite projective geometry provide the construction of new minimal codes (used in particular in secret sharing schemes, to model the access structures) and give rise to multiple of s-ary (where s is the power of 2 to the r and r being a divisor of m) simplex linear codes (whose the dual are the perfect s-ary Hamming codes) over an extension field.

3.16 Gene Prioritization and Rank Aggregation

Olgica Milenkovic (University of Illinois – Urbana Champaign, US)

Joint work of Milenkovic, Olgica; Farnoud, Farzad; Kim, Minji; Raisali, Fardad

We consider the problem of ranking genes according to their likelihood of being implicated in the onset and progression of a disease. Rankings of this form are known as gene prioritizations, and they are used to govern experimental knockout tests. Many software tools exist for prioritizing genes using similarity criteria with respect to genes already known to be involved in the disease, termed disease training genes. Similarity criteria may be as varied as sequence similarity, transcription factor binding cites, expression, and annotation. All individual rankings in one-to-one correspondence with the similarity criteria are aggregated via order statistics methods, which rely on multiple null hypothesis that are hard to test or even potentially accurate.

We propose analyzing the problem via a new combinatorial and information- theoretic approach, based on distance based aggregations. The distances used represent novel extensions of the Kendall distance, used to measure swap distance between two permutations, and the Bregman-Lovasz distance, used to measure distances between rankings and ratings. We also describe integer programming relaxations for the aggregation problems with positional relevance constraints. The distance-based methods outperform order statistics methods in almost all performed tests.

3.17 Neuroscience-inspired Network Decoder

Katherine Morrison (University of Northern Colorado, US)

Joint work of Morrison, Katherine; Curto, Carina

When the brain encounters the same stimulus presented multiple times, different initial neural responses will typically be observed, and yet the brain is still able to determine what stimulus it has encountered. Thus, the brain must be performing some form of error correction. We present a biologically plausible mechanism by which this error correction may be performed. In particular, we show that this decoding algorithm can perform pattern completion, which was previously believed impossible with network models of this form.

3.18 On the Skew Complexity of Sequences with Applications to Algebraic Decoding

Vladimir Sidorenko (Universität Ulm, DE)

Joint work of Sidorenko, Vladimir; Schmidt, Georg; Wachter, Antonia; Bossert, Martin; Li, Wenhui

Linear complexity of a sequence over a field can be efficiently computed using the Berlekamp-Massey algorithm. It will be shown how linear complexity of multiple sequences, which can described using the language of polynomials over a field, can be generalized to skew complexity for the case of skew (or linearized or twisted) polynomials. We show how the skew complexity can be applied for an efficient algebraic decoding of interleaved Reed-Solomon codes and interleaved Gabidulin codes.

3.19 Optimal Index Codes with Near-extreme Rates

Vitaly Skachek (University of Tartu, EE)

The min-rank of a digraph was shown by Bar-Yossef et al. to represent the length of an optimal scalar linear solution of the corresponding instance of the Index Coding with Side Information (ICSI) problem. In this work, the graphs and digraphs of near-extreme min-ranks are studied. Those graphs and digraphs correspond to the ICSI instances having near-extreme transmission rates when using optimal scalar linear index codes. In particular, it is shown that the decision problem whether a digraph has min-rank two is NP-complete. By contrast, the same question for graphs can be answered in polynomial time.

3.20 On Multiply Constant Weight Codes

Patrick Sole (Télécom Paris Tech, FR)

Joint work of Sole, Patrick; Chee, Yeow Meng; Cherif, Zouha; Danger, Jean-Luc; Guilley, Sylvain; Kiah, Han Mao; Kim, Jon-Lark; Zhang, Xiande

Motivated by the security of embarked system a generalization of constant weight codes is introduced and studied. Construction based on parallelism of designs, concatenation of codes, and pseudo-product codes are derived. The asymptotic performance problem is solved completely.

3.21 Index Coding: Fundamentals, Applications, and Recent Progress

Alex Sprintson (Texas A&M University – College Station, US)

Index Coding is one of the central problems in wireless network coding. It has deep connections to many fundamental problems, such as coloring, determining network coding capacity, and interference alignment. This area has been the subject of intensive research and many new insights have been gained over recent years. In this talk, we will survey the fundamentals of index coding, including algorithms for code design, the complexity of such algorithms, and lower and upper bounds on the capacity region of index coding instances. We will also discuss the equivalence between index coding and interference alignment and network coding, as well as the implications of this equivalence. We will also discuss open problems and directions for future research.

3.22 List Decoding of Subspace Codes

Anna-Lena Trautmann (Universität Zürich, CH)

Joint work of Trautmann, Anna-Lena; Rosenthal, Joachim; Silberstein, Natalia

The finite Grassmannian $\mathcal{G}_q(k,n)$ is defined as the set of all k-dimensional subspaces of the ambient space $\mathcal{F}_q^n$. Subsets of the finite Grassmannian are called constant dimension codes and have recently found an application in random network coding. In this setting codewords from $\mathcal{G}_q(k,n)$ are sent through a network channel and, since errors may occur during transmission, the received words can possible lie in $\mathcal{G}_q(k',n)$, where $k' \neq k$.

In this talk, we study the balls in $\mathcal{G}_q(k,n)$ with center that is in $\mathcal{G}_q(k,n)$, for simplicity. We describe the balls with respect to the subspace metric. Moreover, we use two different techniques for describing these balls, one is the Plücker embedding of $\mathcal{G}_q(k,n)$, and the second one is a rank decomposition of the matrix representation of the codewords.

With these results, we consider the problem of list decoding a certain family of constant dimension codes, called lifted Gabidulin codes. We describe a way of representing these codes by linear equations in either the matrix representation or a subset of the Plücker coordinates.

The union of these equations and the equations which arise from the description of the ball of a given radius in the Grassmannian describe the list of codewords with distance less than or equal to the given radius from the received word.

3.23 Trapping Set Structure of Minimum Weight Codewords in Regular LDPC Codes

Bane Vasic (University of Arizona – Tucson, US)

Joint work of Vasic, Bane; Khatami, Seyed Mehrdad; Danjen, Ludovic; Nguyen, Dung V.

We present an efficient algorithm for finding all low-weight codewords in a given quasi-cyclic (QC) low-density parity-check (LDPC) code with a fixed column-weight and girth. A low-weight codeword is viewed as an (a; 0) trapping set, and topologically different (a; 0) trapping set are obtained from smaller trapping sets. The method can be used to construct QC codes with given minimum distance and to determine the multiplicity of the low-weight codewords with different trapping set structure.

3.24 Coding for Combined Block-symbol Error Correction

Pascal Vontobel (HP Labs – Paolo Alto, US)

Joint work of Vontobel, Pascal; Roth, Ron M
Main reference R. M. Roth, P. O. Vontobel, "Coding for Combined Block-Symbol Error Correction," arXiv:1302.1931v1 [cs.IT], 2013.
URL http://arxiv.org/abs/1302.1931v1

Many data transmission and storage systems suffer from different types of errors at the same time. For example, in some data storage systems the state of a memory cell might be altered by an alpha particle that hits this memory cell. On the other hand, an entire block of memory cells might become unreliable because of hardware wear-out.

Such data transmission and storage systems can be modeled by channels that introduce symbol errors and block (i.e., phased burst) errors, where block errors encompass several contiguous symbols. Moreover, if some side information is available, say based on previously observed erroneous behavior of a single or of multiple memory cells, this can be modeled as symbol erasures and block erasures.

We present novel error correction coding schemes that can deal with certain setups that include both symbol and block errors and both symbol and block erasures. At the end of the talk, we will pose some open problem w.r.t. the existence of efficient decoders for certain scenarios.

3.25 On Network Codes and Partial Spreads

Wolfgang Willems (Universität Magdeburg, DE)

Let $\mathcal{G}_q(n, \ell)$ denote the Grassmannian (subspaces of dimension ℓ of an n-dimensional vector space over F_q). We discuss bounds of the function

$$A_q(n, d, \ell) = \max\{|C| : C \subset \mathcal{G}_q(n, \ell), d(C) \geq d\}$$

where the distance is the subspace distance. We mainly focus on the case $d = 2\ell$, i.e., on the maximal size of a "partial ℓ-spread" in an n-dimensional space. Moreover, we answer a question posed by Bu in the seventieth and improve an upper bound of Etzion and Vardy.

Participants

- Daniel Augot
Ecole Polytechnique – Palaiseau & INRIA, FR
- Angela Barbero
University of Valladolid, ES
- Alexander Barg
University of Maryland – College Park, US
- Eimear Byrne
University College Dublin, IE
- Pascale Charpin
INRIA, FR
- Gerard Cohen
ENST – Paris, FR
- Stark C. Draper
University of Toronto, CA
- Iwan M. Duursma
University of Illinois – Urbana Champaign, US
- Salim El Rouayheb
Illinois Inst. of Technology, US
- Marcelo Firer
State University of Campinas – Brazil, BR
- Heide Gluesing-Luerssen
University of Kentucky, US
- Elisa Gorla
Université de Neuchâtel, CH
- Marcus Greferath
University College Dublin, IE
- Hamed S. Hassani
EPFL – Lausanne, CH
- Michael Heindlmaier
TU München, DE
- Tor Helleseth
University of Bergen, NO
- Werner Henkel
Jacobs University – Bremen, DE
- Tracey Ho
CalTech – Pasadena, US
- Tom Høholdt
Technical Univ. of Denmark, DK
- Jørn Justesen
Technical Univ. of Denmark, DK
- Axel Kohnert
Universität Bayreuth, DE
- Margreta Kuijper
The University of Melbourne, AU
- P. Vijay Kumar
IISc – Bangalore, IN
- Michael Lentmaier
Lund University, SE
- Hans-Andrea Loeliger
ETH Zürich, CH
- Felice Manganiello
Clemson University – South Carolina, US
- Muriel Medard
MIT, US
- Sihem Mesnager
University of Paris VIII, FR
- Olgica Milenkovic
University of Illinois – Urbana Champaign, US
- Katherine Morrison
Univ. of Northern Colorado, US
- Joachim Rosenthal
Universität Zürich, CH
- Vladimir Sidorenko
Universität Ulm, DE
- Vitaly Skachek
University of Tartu, EE
- Roxana Smarandache
University of Notre Dame, US
- Patrick Solé
Télécom Paris Tech, FR
- Emina Soljanin
Bell Labs – Murray Hill, US
- Alex Sprintson
Texas A&M University – College Station, US
- Vladimir D. Tonchev
Michigan Tech University – Houghton, US
- Anna-Lena Trautmann
Universität Zürich, CH
- Bane Vasic
Univ. of Arizona – Tucson, US
- Pascal Vontobel
HP Labs – Paolo Alto, US
- Judy L. Walker
Univ. of Nebraska – Lincoln, US
- Wolfgang Willems
Universität Magdeburg, DE
- Oyvind Ytrehus
University of Bergen, NO
- Jiun-Hung Yu
ETH Zürich, CH

Report from Dagstuhl Seminar 13352

Interaction with Information for Visual Reasoning

Edited by

David S. Ebert[1], Brian D. Fisher[2], and Petra Isenberg[3]

1 **Purdue University, US,** ebertd@purdue.edu
2 **Simon Fraser University – Surrey, CA,** bfisher@sfu.ca
3 **INRIA Saclay – Île-de-France – Orsay, FR,** petra.isenberg@inria.fr

Abstract

From August 26–August 30, 2013 Seminar 13352 was held at Dagstuhl on the topic of "Interaction with Information for Visual Reasoning." The seminar brought together a group of cognitive scientists, psychologists, and computer scientists in the area of scientific visualization, information visualization, and visual analytics who were carefully selected for their theoretical and methodological capabilities and history of interdisciplinary collaboration. During the seminar seven discussion groups were formed during which the role of interaction for visualization was carefully reflected on. We discussed in particular the value, structure, and different types of interaction but also how to evaluate visualization and the idea of 'narrative' as applied to visual analytics. This report documents the program and short summaries of the discussion groups for the seminar.

Seminar 26.–30. August, 2013 – www.dagstuhl.de/13352
1998 ACM Subject Classification H.5.2 [Information Interfaces and Presentation]: User Interfaces – Graphical User Interfaces (GUI), I.3.6 [Computer Graphics]: Methodology and Techniques – Interaction Techniques
Keywords and phrases Interaction, visualization, visual analytics, cognitive science, psychology
Digital Object Identifier 10.4230/DagRep.3.8.151

1 Executive Summary

David S. Ebert
Brian D. Fisher
Petra Isenberg

Scientific and information visualization researchers routinely build and evaluate interactive visualization systems to aid human reasoning. However, this work is often disconnected from the methodological and theoretical tools developed by the cognitive and social sciences to address the complexities of human thought processes. Those tools and methods can help us to understand human perception and understanding of data visualization, but typically do not address how rich interaction with computational processes could be engineered to support better decision-making. Yet, an increasing number of researchers are turning to the question of how to best engineer interaction techniques for visualization and how to best study and understand their influence on cognition, insight formation, and also efficiency and effectiveness of work. The goal of this seminar was to bring together researchers in cognitive science and psychology with researchers in the field of visualization to discuss the value that interaction can bring to visualization, how best to study it, and how research on interaction in cognitive science can be best integrated into visualization tools and systems to the benefit of domain experts or casual users of these tools.

2 Table of Contents

3 About the Seminar

Participants

The seminar brought together a diverse group of international cognitive scientists, psychologists, and computer scientists in the area of scientific visualization, information visualization, visual analytics, and human computer interaction. All participants (see Figure 1) were carefully selected for their theoretical and methodological capabilities and history of interdisciplinary collaboration. Thirty participants joined the seminar, out of which seven had a background in psychology and the remainder were primarily computer scientists in training. Eleven participants were female and three in total came from industry. For about one third of participants this seminar was their first Dagstuhl event. Figure 2 shows gender balance and country statistics for all participants.

Format

The seminar followed a format largely based on breakout groups. The first day of the seminar involved short introduction slides for each participant with longer 15 minute invited talks from the domains of cognitive science and psychology. Tuesday the first four breakout groups discussed topics on mixed initiative interaction, crowd interaction, the value of interaction, and conceptual structures of interaction. Each breakout group was comprised of participants with mixed backgrounds in computer science and cognitive science/psychology and had the goal to work on a specific problem related to the title of the breakout group. Wednesday, each participant switched to a second breakout group on topics: evaluation of interaction, magical interaction, and visual narrative. Almost all participants then went together on a social event to visit the Völklinger Hütte (see Figure 3). Thursday morning we heard a presentation about DBLP, and continued with the second breakout group sessions. Thursday ended with discussion in the breakout groups about publishable results from the seminar and working towards establishing a publication plan of action. These were then presented in front of the whole group. Two invited talks started the day on Friday which ended with a discussion on publication venues that would be beneficial for both the computer scientists and cognitive scientists/psychologist. Table 1 gives an overview of the seminar schedule and the following list includes the titles of the individual breakout groups.

- Mixed Initiative Interaction
- Crowd Interaction in Visual Reasoning
- Magical Interaction
- Evaluating Interaction for Visual Reasoning
- Conceptual Structures of Interaction for Visual Reasoning
- Visual Narrative
- The Landscape of Explanations for the Value of Interaction for Visual Reasoning

Output

The organizers and participants planned to publish results from the breakout groups as a Morgan Claypool mini series on interaction for which David Ebert is the series editor. Working groups have been invited to publish their results there. Furthermore we are in contact with an international journal for an open-call special issue to further push the importance of interaction for visual reasoning as an emerging topic in the domain of visualization.

Figure 1 Group picture of all participants taken in the sun outside the Dagstuhl chapel.

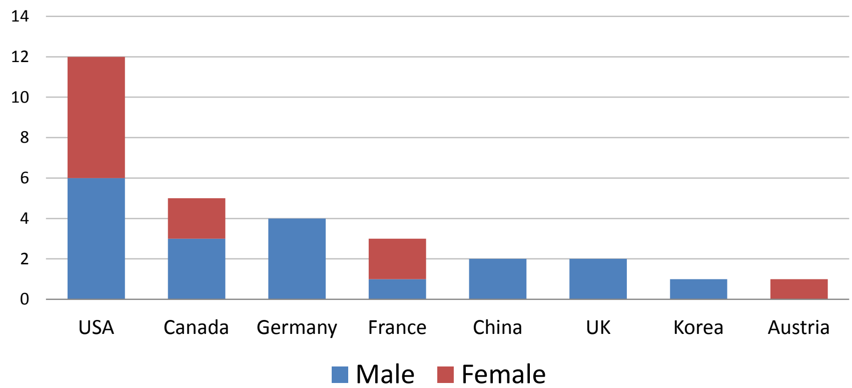


Figure 2 Gender balance and country of participants' home institution.

Figure 3 Excursion to Völklinger Hütte allowed for less structured research conversations and an interesting lesson about the ironworks (picture courtesy of Fanny Chevalier).

Table 1 Schedule of the seminar. Details on talks and breakout groups follows further below.

Monday	*Tuesday*	*Wednesday*	*Thursday*	*Friday*
Introduction to the Seminar Introduction Participants I	Discussion: Breakout Groups	Discussion: Breakout Groups II	Presentation: DBLP	Invited Talks: Sketching
Introduction Participants II	Breakout Groups I	Breakout Groups II	Breakout Groups II	Discussion on Publication Venus & Closing
Introduction Participants III	Breakout Groups I	Social Event: Völklinger Hütte	Discussion Writing Groups I	
Invited Talks Psychology	Report from Breakout Groups I	Social Event: Völklinger Hütte	Discussion Writing Groups II	

The Dagstuhl team performed an evaluation at the end of the seminar. The responses were primarily positive with the overall quality of the seminar rated with a 10/11. All 16 respondents reported that they agreed or agreed completely that the seminar inspired new ideas for their own work, development or teaching. 15/16 respondents agreed that the seminar inspired joint work and all 16 respondents said that the seminar led to insights from neighboring fields. This is exciting as it shows that the seminar's goal of bridging the gap between two research communities was successfully met. In terms of improvements, several participants suggested to leave more room for (impromptu) talks and that it would have helped to prepare participants more prior to arriving at Dagstuhl. We take these as suggestions for the organization of possible future seminars.

Acknowledgments

We would like to acknowledge the valuable input of Dr. Shixia Liu (Microsoft Research Asia) in the application and preparation of this seminar.

4 Overview of Talks

4.1 Interacting with Information – Overview of Past & Current Work

Simon Attfield (Middlesex University, GB)

Joint work of Attfield, Simon; Okoro, Efeosasere

In this talk I presented myself and my research interests. I discussed how an interest in information interaction focusing on domain areas such as journalism, e-discovery, intelligence analysis and healthcare has led to an interest in sensemaking and how it can be supported through digital design. Something I am currently interested in is information structuring during sensemaking, such as howit can be characterised and the comparative effects of different kinds of structuring, such as narrative and argumentation. For both of these questions I am working with Efeosasere Okoro to develop a relational language for capturing the semantics of user generated information structures and looking at the comparative effects of different structuring conventions on task performance anduser experience.

4.2 Cognitive Science of Representational Systems

Peter C.-H. Cheng (University of Sussex – Brighton, GB)

In addition to giving a brief overview of Peter Cheng's main areas of research the talk focused on how cognitive science can inform the design of representational systems to support complex problem solving and learning in conceptually rich domains. It seems necessary to combine our understanding of higher cognition, forms of external representations and the nature of knowledge within a Representational Epistemic approach in order to successfully design graphical displays that can fully accommodate the many diverse tasks that are typically found in complex domains. Novel notations and visualisation were briefly presented for event scheduling, personnel rostering, production planning and scheduling, high school algebra, electricity, particle mechanics, probability theory, propositional logic and syllogisms.

4.3 Flexible Perception of Structure in Viz & Education

Steve Franconeri (Northwestern University – Evanston, US)

Selective attention allows us to filter visual information, amplifying what is relevant and suppressing what competes. But recent work in our lab suggests another role – flexibly extracting and manipulating visual structure. Selective attention allows us to group objects with similar features, extract spatial relationships between objects, and imagine manipulations of objects. An understanding of these mechanisms has concrete implications for the design of visualizations across science and education.

4.4 A Sample of Sketching Research in Cognitive Science

Steve Franconeri (Northwestern University – Evanston, US)

Sketching is a tool for visual thinking. It helps people explore information sets that are too large to hold (or process) in working memory. It helps people re-organize information, allowing it to be seen from a new perspective. I presented two case studies of the power of sketching – in the first, sketching reveals how people understand a problem within an image, and in the second, sketching facilitates insight by promoting for aging through visual relationships.

4.5 Human Interactions in Abstract Visual Spaces

Wayne D. Gray (Rensselaer Polytechnic Institute, US)

Human behavior is interactive behavior. Behavior emerges from the interaction of bounded cognition with the natural or designed task environment and task goals. Topics covered included: (a) Interactive routines, (b) The eye-hand span, (c) Analogies for memory, (d) Local, not global optimization, (e) Modeling the whole human (not just the convenient bits!), (f) Modeling pre-attentive and attentive visual processes, (g) Tools for the statistical analysis of visual saliency and similarity, (h) Tools for analyzing eye data, and (i) Possibilities for more indepth talks including *The Cognitive Science of Natural Interaction* and *Elements of Extreme Expertise*.

4.6 Interacting with Visual Representatives

David Kirsh (University of California – San Diego, US)

Topics covered were: a) what is the difference between epistemic and pragmatic actions – and why they matter for visualization theory; b) explain a core(epistemic) interactive strategy used in reasoning with visualizations. This involves mentally projecting a structure onto an external structure, for instance, an illustration, a geometric figure, a manipulable visualization, then physically realizing that mentally projected structure by altering the external structure or visualization, then start this a) perceive mental project structure; b) create structure process over again by mentally projecting onto this newly altered structure; c) explain the importance of the difference between explicit and implicit encoding of information: when does representation A encode information more explicitly than representation B – good visualizations encode the right information more explicitly; d) what does interaction do for us that we could not do as well in our heads?

4.7 Sketching and Embodied Cognition

David Kirsh (University of California – San Diego, US)

I explore two topics that highlight the power of embodied cognition: using the process of making a sketch to help decide whether a stone is a genuine tool shaped by prehistoric humans as opposed to a visually similar stone that occurs naturally. This shows how the process of physically working with a sketch or illustration teaches us more than just looking at the same illustration or sketch. The second topic introduces another instance of 'modeling or simulation to drive cognition.' In dance there is a process called marking that has many of the same virtues of sketching but this time using the body instead of paper and pencil. Marking is a process in dance practice where a dances sketches a phrase, using less energy and working on aspects of that phrase rather than performing the whole phrase in all its complexity and with all the effort and speed required for full out performance. Our study of marking showed that marking can help a person explore aspects of a movement, one by one; it also allows a person to bring certain elements or dimensions of the phrase into focus. This idea that modeling or imperfectly simulating a complex process can lead to insight that is hard or impossible to obtain by studying the fullprocess is found in practices other than dance and lithic sketching. In painting artists make practice sketches in order to explore elements they want to highlight or get just right. By making simplified models they bring aspects or elements or dimensions of the complex thing they want to paint into better focus, ensuring they capture features they might otherwise have missed. Attached is a paper published on marking and other issues related to interactivity "Embodied Cognition and the Magical Future of Interaction design."

4.8 Toward Systematic Design of Different Interactive Visualization Components

Kamran Sedig (University of Western Ontario, CA)

Visualization tools can support and enhance the performance of cognitive activities such as sense making, problem solving, and analytical reasoning. To do so effectively, however, a human-centered approach to their design and evaluation is required. This presentation highlights a number of different issues that we have been investigating in order to identify some of the main components of interactive visualizations in order to develop a systematic approach for their design and analysis. A few of the issues that are discussed include: Interaction, interactivity, cognitive activities (i. e., reasoning), and visual representations. A recent paper dealing with interaction design for visual representations is also uploaded.

5 Working Groups

5.1 Visual Narratives

Simon Attfield (Middlesex University, GB), Jörn Kohlhammer (Fraunhofer IGD – Darmstadt, DE), Catherine Plaisant (University of Maryland – College Park, US), Margit Pohl (TU Wien, AT), Huamin Qu (The Hong Kong University of Science & Technology, HK), Michelle X. Zhou (IBM Almaden Center – San José, US)

As a group we explored the application of the notion of 'narrative' to visual analytics. We defined narrative as, 'A sequence of events connected in a meaningful structure where the connecting principles are time, causation, logic, rationale and/or entity relationships.' We considered that a 'visual narrative' is the visual rendering of this, and that an 'Interactive visual narrative' is a visual narrative where interaction occurs during construction and/or presentation. For the construction phase some key considerations are element selection and editing and annotating the narrative. For presentation some key considerations are pace, branching, overview and detail, and annotation.

We agreed that narrative can bring structure to information and that this can support comprehension, recall, personal and public audit of an analysis, and also help support analytic systematicity and influence decisions.

We agreed on at least four kinds of narrative as applied to visual analytics. These are:

Stories in the data – Stories which recount temporal and/or causal sequences within data. This is relevant where the data has a temporal dimension or when time is used to 'unfold' data. It is a selective and possibly interpreted account of what the data 'says' structured around temporal associations. It may provide a 'natural' way of thinking about data when it comes to identifying particular kinds of pattern, generating causal explanations, making predictions of the future and for supporting higher-level categorisation (e. g. determining intent of actors during legal cases for crime classification).

The stories of analysis – Stories which communicate the provenance of an analytic outcome by way of a history of the analytic process. It depicts the process through which an analysis was conducted and conclusions reached. Visibility of this story can offer support to the analyst and to others either during or after an analysis for reflecting what was done, auditing and interpreting outcomes. To support construct of this story, data might be gathered during the analysis by automated logging and/or manual annotation. Considerations in abstracting a meaningful narrative include: What are the meaningful agents and objects? What are the meaningful units of activity? How do you detect and represent analytic branching? How do you make a narrative engaging and build in 'Dramaturgie' (crisis/outcome, tension/release)? How do you map from low-level events (as captured by logging) to a meaningful story of progressive sensemaking?

The 'logic' of the conclusion – This is a story which recounts the (probably informal) logic of an analytic outcome. It links elements in terms of premises (both observed and assumed) and conclusions into a coherent argument. It is different from a 'story in the data' and a 'story of the analysis' in that it specifically links elements in terms of the way that one proposition supports another, rather than how one event led to another. It is a sequence of ideas akin to a logical or mathematical proof in which the relations between ideas are determined by relations of implication and not the chronology in which they occurred. An advantage of this kind of story is that its form can make explicit assumptions which might otherwise remain implicit.

Educational narratives – A story generated for pedagogical motives intended for teaching others how to perform or interpret an analysis or how to use complex analytic tools. A key role of story in this case is to lead the student through a series of ideas which may progressively build in a way that is engaging, accessible and memorable. Where the learning supports informed decision making, for example in the case of educating patients, there may be considerations of how presentational elements, such as order, might affect bias.

Finally, we discussed interaction issues as these relate to narrative in visual analytics. We identified three phases of user involvement which deserve consideration. They are: **Data gathering**, raising questions of how this is done; **Narrative construction**, raising questions of how this is done and when; and **Presentation**, raising questions about the provision of overviews for the sake of coherent user mental models, the level of user involvement in interaction (lean forward vs lean back), the pace of presentation, granularity and abstraction of the narrative, and the tension between telling an engaging and representative story.

5.2 Evaluating Interaction for Visual Reasoning

Anastasia Bezerianos (University Paris South, FR), Mary Czerwinski (Microsoft Research – Redmond, US), Brian D. Fisher (Simon Fraser University – Surrey, CA), Steve Franconeri (Northwestern University – Evanston, US), Wayne Gray (Rensselaer Polytechnic Institute, US), Petra Isenberg (INRIA – Saclay, FR), Bongshin Lee (Microsoft Research – Redmond, US), Jinwook Seo (Seoul National University, KR)

One of the main challenges in designing interactive visual analytics systems is to measure the effectiveness of their interaction designs. Traditional performance measures such as task completion time and error rate often fail to demonstrate the value of interacting with visual representations and the effect of interaction on the analysts' reasoning process. Measuring insights is also limited in that it is hard to replicate and quantify, and it does not capture the role of the interactive system in the process that leads to the insightful moments. The role of interaction is thus difficult to tease out with an insight based evaluation. The goal of this paper is to present to the visual analytics community alternative measures related to interaction, human reasoning and analysis processes, borrowed and adapted from the field of cognitive psychology. The article that will be written based on the discussion at Dagstuhl will be structured around the design of a study to evaluate different interaction modalities for visual reasoning. It will discuss both high level questions, such as the formation and evolution of a research question, and low level aspects including the choice of evaluation tasks and methodologies. This article will serve both as an introduction to alternative evaluation measures and methodologies adapted from cognitive psychology, as well as a walkthrough example for researchers on how to formalize research hypotheses and structure evaluations around them.

5.3 The Landscape of Explanations for the Value of Interaction for Visual Reasoning

Sheelagh Carpendale (University of Calgary, CA), Anastasia Bezerianos (University Paris South, FR), Peter Cheng (University of Sussex – Brighton, GB), Brian D. Fisher (Simon Fraser University – Surrey, CA), Steve Franconeri (Northwestern University – Evanston, US), Daniel Keefe (University of Minnesota – Twin Cities, US), Bongshin Lee (Microsoft Research – Redmond, US), Chris North (Virginia Polytechnic Institute – Blacksburg, US)

Our intuition tells us that interaction is a really important factor for visual reasoning. As a community we have generated a wealth of examples of interaction techniques. However, there is little consensus about how to explain how interactions actually create these benefits. By examining a series of diverse interaction techniques in light of visual reasoning we will begin mapping the landscape of explanations of how these interactions add value in terms of visual reasoning. We will use a semi-structure knowledge acquisition process to gather a series of interaction examples that each have three parts: a short stop motion storyboard, an accompanying verbal explanation of the task and system; and, in particular, the creators explanations of the perceived benefits in terms of visual reasoning. By intentionally both sampling for diversity and analyzing from the perspectives of multiple research disciplines we hope to add richness to this discussion. The contribution of this paper is an exploration the space of alternative explanations to expand our understanding of the value of interaction for visual reasoning.

5.4 Mixed Initative Interaction

Christopher Collins (University of Ontario, CA), Simon Attfield (Middlesex University, GB), Fanny Chevalier (University of Toronto, CA), Mary Czerwinski (Microsoft Research – Redmond, US), Heidi Lam (Google Inc. – Mountain View, US), Catherine Plaisant (University of Maryland – College Park, US), Christian Tominski (Universität Rostock, DE), Michelle X. Zhou (IBM Almaden Center – San José, US)

The group defined mixed initiative interaction as a type of interaction for visual reasoning in which the human analyst and the visualization system both are active participants in the interaction. In a traditional interaction scenario, the visualization software is reactive, responding to inputs from the analyst. In MI interaction, the system would play a more active role, for example, making suggestions about appropriate views or next steps in the analysis process. There are then two directions of interaction: human to system, e.g. applying filters, making selections, loading new data; system to human, e.g. suggesting views, suggesting next steps, automatic highlighting of potentially interesting part of a view.

Mixed initiative interaction has been studied for several years, but remains on the periphery of mainstream visualization research. Systems in this area are often called 'smart visualization' or 'intelligent user interfaces'. However, it seems the community is skeptical due to the cost of error: if a smart visualization system suggests a particular representation type, or an analysis process which is inappropriate to the data or current task requirements, then an analyst could become frustrated, or, worse, may come to incorrect conclusions, biased by the underlying interaction model. Other challenges in this research include being able to gather appropriate and sufficient user data to create a model of the user, such as understanding their level of experience, preferences, prior domain knowledge, etc. As this sort of data is difficult to gather and often inconclusive, we focussed our discussion on MI interaction possibilities in scenarios where we do not have prior knowledge about the analyst. Our discussions lead to a list of factors which can be used to evaluate the success of system-initiated interaction prompts:

- Are they timely? Are suggestions provided at the right time or do they interrupt the analyst's flow?
- Does the system take initiative sparingly? If the system takes the initiative too often, the analyst may become fatigued and ignore suggestions.
- Are system suggestions appropriate? Is the system suggesting views, prompts, or other cues which enhance the analysis experience and potential for insight? Or do they lead to incorrect conclusions about the data?
- Is the provenance of system suggestions transparent? Can the human analyst understand why the system makes any given suggestion?

Scenarios where MI interaction may be useful include a new analyst using a system for the first time and requiring tutorial-style guidance. In this scenario, the system may not know much about the characteristics, prior knowledge, interaction styles and preferences of the analyst and has to provide assistance based on characteristics of the data and the current interaction session, and perhaps a crowd-driven model of the way other analysts have used the system.

Human to System Interaction (Human Initiative)

We called the types of traditional interactions, such as selecting data items, panning and zooming a view, "*explicit* interactions". Newer forms of interaction, such as 'model steering' by repositioning items in a visualization to indicate prior knowledge about their relatedness, are also important inputs to a mixed initiative system.

Where we focused our discussion was on new forms of cues which may be gathered by the system in a mixed initiative interaction model to improve the quality and timeliness of prompts and suggestions. We called these inputs 'implicit interactions'. We enumerated the following list of potential 'implicit interactions' which could be tracked by a visualization system and used to decide when and how to take initiative:

- Dwell time (eye gaze, touch, or mouse cursor)
- Facial gestures
- Highlighting / copying behaviour
- Repeated actions
- Body position / gestures (proxemics)
- Thrashing—changing actions / direction
- Emotional indicators
- Physiological indicators
- Mouse signatures
- Keyboard signatures
- Repositioning items on the screen
- The history of what they have explored already (the analysis process)

These implicit interactions could provide a wealth of data to a mixed initiative system, but would also have drawbacks which need to be investigated, including privacy concerns, potential for reinforcing actions (encouraging 'tunnel vision'), or ambiguity of the meaning of the indicators. For example, physiological and behavioural indicators of excitement and annoyance may be quite similar. Which implicit interactions would be most important and how they could work together to create a profile of the analyst state are areas of future research inspired by our discussions.

System to Human Interaction (System Initiative)

Others have researched system-initiated interaction driven by user profiles and data characteristics, so we targeted our discussions on the types of feedback a system could provide based on analysis of implicit interaction data. The timing of system-initiated interaction is crucial: ideally it is timely, does not interrupt the flow of analysis and human-initiated interaction, and is appropriate to the data and task. Design decisions to consider in future work include: (a) how to present interaction and analysis suggestions, (b) how to reveal the provenance of guidance (why a suggestion is made by the system), (c) how to encode the confidence level the system has in a suggestion, etc. We recommended that system feedback could also be subtle or implicit. For example, if the system senses the analyst is “lost” or “stressed”, rather than asking “are you stressed?” it could simply provide additional on screen help, adjust or simplify the interface, and suggest alternative views.

We explored a variety of system responses which may be appropriate if the indicators strongly point to the need for system-initiated interaction. Specifically, we looked at possible responses to various detected emotional states, such as frustration, confusion, boredom, interest, and engagement. System responses may include: show more views like the current view, show views different from the current view, show what other people (all people / people like me / experts) did in similar situations, offer help, or simplify the view (remove a data dimension or perform aggregation). MI interaction should be flexible and perhaps system-initiation should be turned off automatically after the analyst does not acknowledge or use system suggestions over a long period of time.

To conclude our meetings, the group brainstormed about a paper outline reporting on our mixed initiative interaction for visual reasoning ideas, and assigned next steps to the participants.

5.5 Conceptual Structures of Interaction for Visual Reasoning

Kelly Gaither (University of Texas – Austin, US), David S. Ebert (Purdue University, US), Thomas Ertl (Universität Stuttgart, DE), Hans Hagen (TU Kaiserslautern, DE), Petra Isenberg (INRIA Saclay, FR), Tobias Isenberg (INRIA Saclay, FR), Jörn Kohlhammer (Fraunhofer IGD – Darmstadt, DE), Margit Pohl (TU Wien, AT), Kamran Sedig (University of Western Ontario, CA)

Interaction is a fundamental element of successful visualization methods and tools. In visualization, interaction can support many low-level and high-level tasks and goals, can support different representation and interaction intends, and can be realized by different techniques. The specific incarnations of the interaction design, however, are driven by the specific application domain, by the tasks being supported, by the type of data being analyzed, by the specific representations being chosen, by potential limitations of computability, and by the needs and requirements of the users. The question that we aimed to analyze is if we can identify general principles of interactions that bridge different domains and are common among tasks, data types, and representations. Can we formulate or propose a language or schema of interaction that is common for most if not all visualization tools and methods, potentially with different dialects?

5.6 Magic Interactions with Information for Visual Reasoning

Daniel Keefe (University of Minnesota – Twin Cities, US), Sheelagh Carpendale (University of Calgary, CA), Peter Cheng (University of Sussex – Brighton, GB), Fanny Chevalier (University of Toronto, CA), Christopher Collins (University of Ontario, CA), Tobias Isenberg (INRIA Saclay, FR), David Kirsh (University of California – San Diego, US), Heidi Lam (Google Inc. – Mountain View, US), Chris North (Virginia Polytechnic Institute – Blacksburg, US), Kamran Sedig (University of Western Ontario, CA), Christian Tominski (Universität Rostock, DE), Xiaoru Yuan (Peking University, CN)

Today, there is much excitement around the concept of "natural user interfaces." The interest is sparked in part by the widespread availability of multi-touch devices, including smart phones and tablets. However, the trend is not limited to these new commercial devices; a variety of recently developed user interface techniques that enable seemingly more direct ways of interfacing with computers have been dubbed "natural." Will these natural interactions define the future of computing? As user interface designers, and in particular as designers and researchers interesting in supporting users as they reason about super-complex information, we have to ask, is "natural" actually the right target? Do we really want to design natural interactions or do we want something else? How about "supernatural" or even "magical" interactions? Our Dagstuhl working group found that the more we thought about the systems and human-computer interfaces that have most influenced us or impacted our work, the more we recognized that (at least the first few times we used these systems) they all felt magical. Some examples include: (1) clicking and dragging a drawing of a cartoon character who then responds "intelligently" by changing his pose in direct response to the user's input, understanding how to move as if by magic; (2) Browsing video data by clicking directly on characters in the video rather than using a slider; (3) Bumping mobile devices to transfer files; and (4) Selecting 3D point clouds just by drawing a 2D lasso. All of these interactions have "the power of apparently influencing the course of events by using mysterious or supernatural forces" and "a quality that makes something seem removed from everyday life, esp. in a way that gives delight" – two properties taken directly from the definition of the word magic. Grounded in findings from the cognitive science research community, we developed several explanations for when and why "magical interactions" seem to work well, including the notions of a different cognitive cost structure for natural vs. magical interactions, superpower and amplification, context/temporal appropriateness, and working with underspecified and imprecise data or applications. Based on these insights, we call for a new research focus that moves beyond "natural user interfaces" and instead targets magic interactions with information.

5.7 Crowd Interaction in Visual Reasoning

David Kirsh (University of California – San Diego, US), Huamin Qu (The Hong Kong University of Science & Technology, HK), Jinwook Seo (Seoul National University, KR), Xiaoru Yuan (Peking University, CN)

How can we harness the intelligent capacity of crowds to reason visually about a topic? In recent years real-time posting of images, video and tweets by citizens in the midst of a natural disaster has provided better information to first responders than information from helicopters and their own officials on the ground. That new source of local information is just starting. It presents a great opportunity for data mining but also for visual reasoning. There are major challenges. How should we enable *intelligent aggregation* of visual and other qualitative data? How should we make *visual and textual summaries* to communicate ideas? How might we use groups to *preprocess data* to support interactive visualizations?

The basic methods of crowd sourcing in use to date are far removed from our goal of using crowds to create meaningful visualizations. Currently, crowds are used for judging the plausibility of statements, for making value judgments, for giving aggregate views, but as yet not for creating visual narratives, intelligent visualizations, or preprocessing qualitative data.

In our group we identified problems that need to be solved to harness crowd power for visual reasoning and qualitative argumentation. Beside aggregation and intelligent summarizing we considered questions of coordination: how should the identification and disbursement of tasks be intelligently managed? Can this be done automatically or can we use Turkers to make themselves into a self-organizing system? How can Turkers be unleashed on a problem and generate requests that when answered in the right way are amenable to visualizing complex problems?

Participants

- Simon Attfield
Middlesex University, GB
- Anastasia Bezerianos
University Paris South, FR
- Sheelagh Carpendale
University of Calgary, CA
- Peter C.-H. Cheng
Univ. of Sussex – Brighton, GB
- Fanny Chevalier
University of Toronto, CA
- Christopher Collins
University of Ontario, CA
- Mary Czerwinski
Microsoft Res. – Redmond, US
- David S. Ebert
Purdue University, US
- Thomas Ertl
Universität Stuttgart, DE
- Brian D. Fisher
Simon Fraser Univ. – Surrey, CA
- Steve Franconeri
Northwestern University – Evanston, US
- Kelly Gaither
University of Texas – Austin, US
- Wayne D. Gray
Rensselaer Polytechnic, US
- Hans Hagen
TU Kaiserslautern, DE
- Petra Isenberg
INRIA Saclay – Île-de-France – Orsay, FR
- Tobias Isenberg
INRIA Saclay – Île-de-France – Orsay, FR
- Daniel Keefe
University of Minnesota – Duluth, US
- David Kirsh
University of California – San Diego, US
- Jörn Kohlhammer
Fraunhofer IGD – Darmstadt, DE
- Heidi Lam
Google Inc. – Mountain View, US
- Bongshin Lee
Microsoft Res. – Redmond, US
- Chris North
Virginia Polytechnic Institute – Blacksburg, US
- Catherine Plaisant
University of Maryland – College Park, US
- Margit Pohl
TU Wien, AT
- Huamin Qu
The Hong Kong University of Science & Technology, HK
- Kamran Sedig
Univ. of Western Ontario, CA
- Jinwook Seo
Seoul Nat. University, KR
- Christian Tominski
Universität Rostock, DE
- Xiaoru Yuan
Peking University, CN
- Michelle X. Zhou
IBM Almaden Center – San José, US