DAGSTUHL
REPORTS

Volume 3, Issue 5, May 2013

Dagstuhl Reports, Vol. 3, Issue 5 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
October, 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.3.5.i

www.dagstuhl.de/dagrep

Report from Dagstuhl Seminar 13192

Tree Transducers and Formal Methods

Edited by

Sebastian Maneth[1] and Helmut Seidl[2]

1 **University of Oxford, GB,** `sebastian.maneth@gmail.com`
2 **TU München, DE,** `seidl@in.tum.de`

Abstract

The aim of this Dagstuhl Seminar was to bring together researchers from various research areas related to the theory and application of tree transducers. Recently, interest in tree transducers has been revived due to surprising new applications in areas such as XML databases, security verification, programming language theory, and linguistics. This seminar therefore aimed to inspire the exchange of theoretical results and information regarding the practical requirements related to tree transducers.

Seminar 5.–8. May, 2013 – www.dagstuhl.de/13192
1998 ACM Subject Classification F.1.1 Models of Computation
Keywords and phrases tree transducers, expressiveness, complexity
Digital Object Identifier 10.4230/DagRep.3.5.1
Edited in cooperation with Marco Kuhlmann

1 Executive Summary

Sebastian Maneth

The Dagstuhl seminar 13192 "Tree Transducers and Formal Methods" was a short two and a half day seminar that took place from May 5th to 8th, 2013. The aim was to bring together researchers from various research areas related to the theory and application of tree transducers. Tree transducers are a classical formalism in computer science, dating back to the early days of compilers and syntax-directed translation. Recently, interest in tree transducers has been revived due to surprising new applications in areas such as XML databases, security verification, programming languages, and linguistics. This seminar was meant to inspire the exchange of theoretical results and practical requirements related to tree transducers. These points were addressed in particular:

- Expressiveness versus Complexity: Which transducers offer the best trade-offs between expressiveness and complexity?
- Implementability under Resource Restrictions: Which transducer models can be executed by devices with bounded resources, e.g., for processing XML streams?
- New Applications: What new challenges do the different application areas of tree transducers raise? What new solutions have been found?
- Open Problems: Which are the most pressing open problems in tree transducer theory?

The seminar fully satisfied our expectations. The 33 participants from 13 countries (Australia, Belgium, Canada, Czech, France, Germany, Great Britain, Hungary, Japan, Poland, Slovakia, Sweden, and the US) had been invited by the organizer Sebastian Maneth to give particular survey talks about their recent research on applications and theory of tree transducers.

There were talks focusing on very practical issues such as Margus Veanes' talk on software verification using symbolic tree transducers (which kicked off the meeting), and also talks on highly challenging theoretical results such as the talk by Emmanuel Filiot on their recent breakthrough of proving that one-wayness of a two-way word automaton is decidable. The other application areas, besides verification, were (1) tree processing (related to databases and search) (2) learning, and (3) linguistics.

The first talk by Veanes on symbolic transducers was followed by Jan Janousek about using pushdown automata to search for tree patterns, in linear order of trees. Symbolic transducers, from a theoretical point of view, were discussed in Heiko Vogler's talk in the afternoon. Input driven pushdown automata, also known as nested word automata or visibly pushdown automata, were discussed with respect to descriptional complexity by Kai Salomaa. The second morning session of the first day was devoted to MSO translations, first about its theory with respect to word and tree translations by Bruno Courcelle, and then concerning a one-pass and linear time implementation model for MSO tree translations: the streaming tree transducer by Loris d'Antoni. The first afternoon section was about higher-order transducers, recursion schemes, and verification, given by Kazuhiro Inaba, Luke Ong, and Naoki Kobayashi. They discussed the open problem of proving context-sensitivity of the unsafe OI-hierarchy, results on model checking of higher-order recursion schemes, and practical approaches to type checking unsafe higher-order tree transducers.

The second day started with theoretical results about word and tree transducers by Emmanuel Filiot and Sebastian Maneth. The latter one was about deciding two database notions, namely determinacy and rewriting, for top–down and MSO tree transducers. Next was a sequence of talks about streaming, by Joachim Niehren, Pavel Labath, and Keisuke Nakano. They discussed practical aspects of early query answering, streaming of macro tree transducers using parallel streams, and stack attributed tree transducers, respectively. Related to streaming was the following talk by Frederic Servais which surveyed recent results on visibly pushdown transducers. The following three talks discussed learning algorithms: first about tree series by Johanna Björklund and Frank Drewes, and then about top–down tree transformations by Adrien Boiret. The last talk of the second day was Florent Jacquemard and Sophie Tison's survey about tree automata with constraints.

The final day started with a talk about natural language processing using transducers, given by Daniel Gildea. It presented applications of multi bottom-up tree transducers to machine translation of natural language. It was followed by a talk by Uwe Mönnich on logical definitions of mildly context-sensitive grammar formalisms. A survey on "the tree-based approach" to natural language grammars was given by Marco Kuhlmann. Damian Niwinski's talk connected to the session of the first day on higher-order schemes: they are equivalent to panic automata, the invention and topic of Damian. An important practical consideration is incremental evaluation: it was discussed for XPath by Henrik Björklund and for succinct regular expressions by Wim Martens.

We thank Schloss Dagstuhl for the professional and inspiring atmosphere it provides. Such an intense research seminar is possible because Dagstuhl so perfectly meets all researchers' needs. For instance, elaborate research discussions in the evening were followed by musical intermezzi of playing piano trios by Beethoven and Mozart, or by table tennis matches and sauna sessions.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 Incremental XPath Evaluation

Henrik Björklund (University of Umeå, SE)

Joint work of Björklund, Henrik; Gelade, Wouter; Martens, Wim

Incremental view maintenance for XPath queries asks to maintain a materialized XPath view over an XML database. It assumes an underlying XML database D and a query Q. One is given a sequence of updates U to D, and the problem is to compute the result of $Q(U(D))$: the result of evaluating query Q on database D after having applied updates U. This article initiates a systematic study of the Boolean version of this problem. In the Boolean version, one only wants to know whether $Q(U(D))$ is empty or not. In order to quickly answer this question, we are allowed to maintain an auxiliary data structure. The complexity of the maintenance algorithms is measured in, (1) the size of the auxiliary data structure, (2) the worst-case time per update needed to compute $Q(U(D))$, and (3) the worst-case time per update needed to bring the auxiliary data structure up to date. We allow three kinds of updates: node insertion, node deletion, and node relabeling. Our main results are that downward XPath queries can be incrementally maintained in time $O(\text{depth}(D)\cdot\text{poly}(|Q|))$ per update and conjunctive forward XPath queries in time $O(\text{depth}(D)\cdot\log(\text{width}(D))\cdot\text{poly}(|Q|))$ per update, where $|Q|$ is the size of the query, and $\text{depth}(D)$ and $\text{width}(D)$ are the nesting depth and maximum number of siblings in database D, respectively. The auxiliary data structures for maintenance are linear in $|D|$ and polynomial in $|Q|$ in all these cases.

References

1 Henrik Björklund, Wouter Gelade, and Wim Martens. Incremental XPath Evaluation. *ACM Transations on Database Systems*, Vol. 35, No. 4, Article 29, November 2010.

3.2 Learning Deterministic Top–Down Tree Transducers with Inspection

Adrien Boiret (Université Lille, FR)

Joint work of Boiret, Adrien; Lemay, Aurélien; Niehren, Joachim

We present a Myhill–Nerode result on deterministic top–down tree transducers, an extension of the preexisting result by Lemay, Maneth, Niehren in 2010.

This result gives us a minimal normal form, and then allows us to devise a learning algorithm on the Gold model (Gold 1978), using behavior examples as an input, and providing the minimal normal form as the output.

3.3 Learning Tree Series

Frank Drewes (University of Umeå, SE)

We give an introduction to the basic ideas of active learning, focusing on the learning of tree languages and tree series from a so-called minimal adequate teacher.

In active learning, the learning algorithm is allowed to request training data by need. The paradigm was designed to improve accuracy and reduce annotation effort, and is particularly appropriate when data is easy to come by, but labelling it is expensive. The labelling resource is typically modelled as an oracle capable of answering certain kinds of queries. The blueprint of such a resource is Angluin's minimal adequate teacher (MAT) which accepts membership queries and equivalence queries [1]. In a membership query, the oracle is asked to label an element as inside our outside the target language L. In an equivalence query, the oracle is given a language model M (e.g., a finite automaton or a grammar) and is expected to return a counter-example to the conjecture that $L(M) = L$, or to acknowledge that L has been correctly acquired.

In this talk, we discuss the inference of tree languages and, more generally, tree series within the MAT model. A tree series is a function from a domain of trees to some algebraic structure, often a semiring or semifield. We focus on the generalisation of Angluin's LSTAR algorithm to trees, and explain central tools and techniques such as observation tables, contradiction backtracking, and proof-of-life contexts by means of an extensive example that covers the main ideas of [2, 3, 4]. The talk concludes with a summary of related results, in which one or more of the components (i) language model, (ii) target class, and (iii) oracle definition has been altered. In doing so, we touch upon the learnability of tree transducers, residual and universal automata, and variations on the classical MAT queries such as correction queries and inclusion queries.

References

1 Dana Angluin. Learning regular sets from queries and counterexamples. *Information and Computation*, 75:87–106, 1987.

2 Frank Drewes and Johanna Högberg. Query learning of regular tree languages: How to avoid dead states. *Theor. Comp. Sys.*, 40(2):163–185, 2007.

3 Frank Drewes and Heiko Vogler. Learning deterministically recognizable tree series. *Journal of Automata, Languages and Combinatorics*, 12:333–354, 2007.

4 Andreas Maletti. Learning deterministically recognizable tree series – revisited. In Symeon Bozapalidis and George Rahonis, editors, *Proceedings of the 2nd international conference on Algebraic informatics*, CAI'07, pages 218–235, Berlin, Heidelberg, 2007. Springer-Verlag.

3.4 From Two-Way to One-Way Finite State Transducers

Emmanuel Filiot (Université Paris 12, FR)

Joint work of Filiot, Emmanuel; Gauwin, Olivier; Reynier, Pierre-Alain; Servais, Frédéric
Main reference E. Filiot, O. Gauwin, P.-A. Reynier, F. Servais, "From Two-Way to One-Way Finite State Transducers," arXiv:1301.5197v2 [cs.FL]. To appear in the Proceedings of IEEE/ACM Logic in Computer Science (LICS), 2013.
URL http://arxiv.org/abs/1301.5197v2

Any two-way finite state automaton is equivalent to some one-way finite state automaton. This well-known result, shown by Rabin and Scott and independently by Shepherdson, states that two-way finite state automata (even non-deterministic) characterize the class of regular languages. It is also known that this result does not extend to finite string transductions: (deterministic) two-way finite state transducers strictly extend the expressive power of (functional) one-way transducers. In particular deterministic two-way transducers capture exactly the class of MSO-transductions of finite strings. In this talk, we address the following definability problem: given a function defined by a two-way finite state transducer, is it definable by a one-way finite state transducer? By extending Rabin and Scott's proof to transductions, we show that this problem is decidable. Our procedure builds a one-way transducer, which is equivalent to the two-way transducer, whenever one exists.

3.5 Forward and Backward Application of Symbolic Tree Transducers

Zoltán Fülöp (University of Szeged, HU)

Joint work of Fülöp, Zoltan; Heiko, Vogler
Main reference Z. Fülöp, H. Vogler, "Forward and Backward Application of Symbolic Tree Transducers," arXiv:1208.5324v1 [cs.FL], 2013.
URL http://arxiv.org/abs/1208.5324

We characterized symbolically recognizable (s-recognizable) tree languages in terms of classical recognizable tree languages and relabelings of infinite range. Also we gave sufficient conditions for that the syntactic composition of two symbolic tree transducers (stt) computes the composition of the tree transformations computed by each stt. We considered forward and backward application of stt and proved that the backward application of an stt to any s-recognizable tree language yields and s-recognizable tree language. We gave a linear stt of which the range is not an s-recognizable tree language. We showed that the forward application of a simple and linear stt preserves s-recognizability.

References

1 Fülöp, Zoltán and Vogler, Heiko. Forward and Backward Application of Symbolic Tree Transducers, http://arxiv.org/abs/1208.5324, 2013

3.6 On the Translations Produced by Multi Bottom-Up Tree Transducers

Daniel Gildea (University of Rochester, US)

Main reference D. Gildea, "On the String Translations Produced by Multi Bottom-Up Tree Transducers," Computational Linguistics, 38(3):673–693, 2012.
URL http://dx.doi.org/10.1162/COLI_a_00108

Multi Bottom-Up Tree Transducers have recently been proposed as a model for machine translation due to the attractive property that they are closed under composition. Tree transducers are defined as relations between trees, but in syntax-based machine translation, we are ultimately concerned with the relations between the strings at the yields of the input and output trees. We examine the formal power of Multi Bottom-Up Tree Transducers from this point of view.

3.7 Higher-Order Tree Transducers and Their Expressive Power

Kazuhiro Inaba (Google Japan, JP)

Joint work of Inaba, Kazuhiro; Maneth, Sebastian
Main reference K. Inaba, S. Maneth, "The Complexity of Tree Transducer Output Languages," in Proc. of the IARCS Annual Conf. on Foundations of Software Technology and Theoretical Computer Science (FSTTCS'08), LIPIcs, Vol. 2, pp. 244–255, Schloss Dagstuhl, 2008.
URL http://dx.doi.org/10.4230/LIPIcs.FSTTCS.2008.1757

This talk reviews and discusses about the expressive power of higher-order tree grammars and transducers. In the literature, two major notions of "high-order" devices have been studied. One that well-known to the transducer community is the line of researches on IO-/OI- hierarchy [1] and high-level transducers [2]. There, trees are the order-0 entities, and order-$(n + 1)$ entities are the functions from order-n entities to an order-n entity. These classes of higher-order hierarchy are known to have beautiful properties. Notably, their expressive power is characterized by n-iterated pushdown stack (i.e., stack of stack of ... of stack), or by n-fold composition of first order macro tree transducers. We show, by fully utilizing the decomposition result, that the languages in OI-hierarchy are context-sensitive [3].

On the other hand, pushed by the recent need for higher-order model checking, broader class of higher-order functions are now investigated [4, 5]. It is called an "unsafe" grammar/transducer (and as a contrast, the former definition is called "safety" restriction), and takes all terms of simply-typed lambda- calculus into account. In particular, it involves a higher-order term containing lower- order free variables, which can never occur under the "safe" construction. Contrary to the safe case, no result is known about the first-order decomposition for unsafe case, nor whether it is included in the class of context-sensitive languages. We discuss the ongoing approach to tackle those open problems.

References

1 W. Damm. The IO- and OI-hierarchies. *Theoretical Computer Science*, 20:95–207, 1982.
2 J. Engelfriet and H. Vogler. High level tree transducers and iterated pushdown tree transducers. *Acta Informatica*, 26:131–192, 1988.

3 K. Inaba and S. Maneth. The complexity of tree transducer output languages. In *Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, pages 244–255, 2008.
4 T. Knapik, D. Niwiński, and P. Urzyczyn. Deciding monadic theories of hyperalgebraic trees. In *Typed Lambda Calculi and Applications (TLCA)*, pages 253–267, 2001.
5 C.-H. L. Ong. On model-checking trees generated by higher-order recursion schemes. In *Logic in Computer Science (LICS)*, pages 81–90, 2006.

3.8 Tree Automata with Constraints: A Brief Survey

Florent Jacquemard (IRCAM & INRIA – Paris, FR)

Joint work of Jacquemard, Florent; Tison, Sophie; Filiot, Emmanuel
URL http://tata.gforge.inria.fr

It is well-known that tree automata define exactly regular languages of trees. However for some problems one sometimes needs to test for equalities and disequalities of subtrees. For instance, ranges of non-linear tree transducers cannot be represented by tree automata. To overcome this problem some extensions of tree automata with tree (dis)equality constraints have been proposed. This talk surveys some of the most important models and their applications. Two families of automata are presented. First, we consider automata with local constraints. They have been used to solve problems related to pattern-matching, tree rewriting and more recently the tree homomorphism problem. Then, motivated by applications to XML processing and security protocols verification, we present a more recent model of tree automata with global constraints.

3.9 Tree Indexing by Deterministic Automata

Jan Janousek (Czech Technical University, CZ)

Joint work of Janousek, Jan; Melichar Borivoj; Flouri, Tomas; Poliak, Martin; Travnicek, Jan

We present a basic survey and the main ideas of tree indexing implemented by deterministic automata. Given a tree of size n, we construct deterministic pushdown automata or deterministic finite tree automata which represent a full index of the tree for subtrees or tree patterns. Given an input tree pattern which matches the tree, its acceptance by the automaton is performed in $\mathcal{O}(m)$ time, where m is the size of the tree pattern, and does not depend on n. We present deterministic pushdown automata which read linear notations of trees and are analogous to deterministic finite automata representing full index of strings: a subtree pushdown automaton is analogous to a string factor automaton. A tree pattern pushdown automaton is then an extension of the subtree pushdown automaton for indexing the tree patterns. Moreover, also oracle versions of these automata can be constructed, as it is in the case for string indexing automata. All these pushdown automata are input-driven and with just one pushdown symbol in their basic versions. Therefore, they are convenient for implementation. Another possibility is to construct a deterministic finite tree automaton representing the index, or a deterministic pushdown automaton reading a tree in a linear notations that corresponds to the finite tree automaton. These possibilities are also discussed and shown.

References

1 Janousek, J., Melichar, B. On Regular Tree Languages and Deterministic Pushdown Automata. In *Acta Informatica*, Vol. 46, No. 7, pp. 533-547, Springer, 2009.

2 Janousek, J. String Suffix Automata and Subtree Pushdown Automata. In: *Proceedings of the Prague Stringology Conference 2009*, pp. 160–172, Czech Technical University in Prague, Prague, 2009.

3 Melichar, B., Janousek, J., Flouri, T. Arbology: Trees and Pushdown Automata. In: *Kybernetika*, vol. 48, No.3, pp. 402-428, 2012.

3.10 Verifying Higher-Order Tree Transducers by Higher-Order Model Checking

Naoki Kobayashi (University of Tokyo, JP)

Joint work of Kobayashi, Naoki;Hiroshi Unno; Naoshi Tabuchi
Main reference N. Kobayashi, H. Unno, N. Tabuchi, “Higher-order multi-parameter tree transducers and recursion schemes for program verification,” in Proc. of the 37th ACM SIGPLAN-SIGACT Symp. on Principles of Programming Languages (POPL’10), pp. 495–508, ACM, 2010.
URL http://dx.doi.org/10.1145/1706299.1706355

We discuss methods for type-checking (unsafe) higher-order tree transducers using forward inference and higher-order model checking. The idea is to represent the forward image as a higher-order recursion scheme, and use higher-order model checking to check the inclusion by the output language. This approach is arguably more efficient in practice, thanks to the recent advance of higher-order model checking algorithms. We present two variations of the method for computing the forward image: one for the combination of regular input languages and a higher-order multi-tree transducer, and the other for the combination of a higher-order input language and a higher-order (single) tree transducer. The former is joint work with Hiroshi Unno and Naoshi Tabuchi, presented at POPL 2010.

3.11 Natural Language Grammars: A Tree-Based Approach

Marco Kuhlmann (Uppsala University, SE)

In this talk I gave three examples of problems in theoretical computational linguistics whose solution, in various ways, has involved the use of tree grammars and tree transducers: the so-called lexicalization of tree adjoining grammars; the definition of a grammar formalism for dependency grammar; and the question about the generative capacity of categorial grammars. I also mentioned some open problems in computational linguistics that may be attacked using tree automata theory.

References

1 Marco Kuhlmann and Giorgio Satta. Tree-Adjoining Grammars Are Not Closed Under Strong Lexicalization. *Computational Linguistics*, 38(3):617–629, 2012.

2 Marco Kuhlmann. Mildly Non-Projective Dependency Grammar. *Computational Linguistics*, 39(2):355–387, 2013.

3 Marco Kuhlmann, Alexander Koller, and Giorgio Satta. The Importance of Rule Restrictions in CCG. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 534–543, Uppsala, Sweden, 2010.

3.12 A Functional Language for Hyperstreaming XSLT

Pavel Labath (University of Bratislava, SK)

Joint work of Labath, Pavel; Niehren, Joachim
Main reference P. Labath, J. Niehren, "A Functional Language for Hyperstreaming XSLT", Research Report, 2013.
URL http://researchers.lille.inria.fr/~niehren/Papers/X-Fun/0.pdf

The problem of how to transform large data trees received on streams with a much smaller memory is still an open challenge despite of a decade of research on XML. Therefore, the current approach of the XSLT working group of the W3C is to provide streaming support only for a small fragment of XSLT 3.0. This has the drawback that many existing XSLT programs need to be rewritten in order to become executable on XML streams, while many others cannot be rewritten at all, since they are defining nonstreamble transformations.

We propose a new hyperstreaming approach that does not require any a priori restrictions. The model of hyperstreaming generalizes on the model of streaming by adding shredding operations for the output stream, so that its parts may be plugged together later on. Many transformations such as flips of document pairs are hyperstreamable but not streamable. We then present the functional language X-Fun for defining transformations between XML data trees, while providing shredding instructions. X-Fun can be understood as an extension of Frisch's XStream language with output shredding, while pattern matching is replaced by tree navigation with XPath expressions.

We also provide a compiler from a fragment of XSLT into X-Fun, which can then be considered as the core of XSLT. We then present a hyperstreaming algorithm for evaluating X-Fun programs which combines a recent XPath evaluator with a traditional functional programming engine. We have implemented a hyperstreaming evaluator for X-Fun and thus for XSLT and compared it experimentally with Saxon's XSLT implementation. It turns out that many XSLT programs become hyperstreamable with good efficiency and without any manual rewriting.

References

1 A. Frisch and K. Nakano. Streaming XML transformation using term rewriting. In *Programming Language Technologies for XML (PLAN-X)*, 2007b.

3.13 Determinacy and Rewriting of Top-Down and MSO Tree Transformations

Sebastian Maneth (University of Oxford, GB)

Joint work of Benedikt, Michael; Engelfriet, Joost; Maneth, Sebastian
Main reference To appear in MFCS'13.

A query is determined by a view, if the result to the query can be reconstructed from the result of the view. We consider the problem of deciding for two given tree transformations, whether one is determined by the other. If the view transformation is induced by a tree transducer that may copy, then determinacy is undecidable, even for identity queries. For a large class of non-copying views, namely compositions of functional extended linear top-down tree transducers with regular look-ahead, we show that determinacy is decidable, where queries are given by deterministic top-down tree transducers with regular look-ahead or by MSO tree transducers. We also show that if a query is determined, then it can be rewritten into a query that works directly over the view and is in the same class as the given query. The proof uses

- uniformizers
- composition results and
- decidability of equivalence.

A uniformizer of a binary relation is a function that for each input of the relation chooses an arbitrary output of the relation. The idea is to construct a representation of the inverse of a view, then to build a uniformizer U for it, and then to build a transducer for the composition of the view, followed by U, followed by the query Q, and to check equivalence of this transducer with they query. The transducers are equivalent if and only if the query is determined by the view. These result will be presented at MFCS'2013 in Vienna.

3.14 Efficient Incremental Evaluation of Succinct Regular Expressions

Wim Martens (Universität Bayreuth, DE)

Joint work of Björklund, Henrik; Martens, Wim; Timm, Thomas

We present a method for efficient incremental evaluation of regular expressions with counters. Such expressions are used in grep, Python, Perl, Ruby, XML Schema, and are being considered for property paths in SPARQL 1.1. Furthermore, they can be exponentially more succinct than "traditional" regular expressions or non-deterministic finite automata as usually studied in the literature. Our evaluation method exploits the counter values in the expressions to avoid an exponential blow-up that the current state-of-the-art algorithms have. In this study, we present a thorough investigation of the use and structure of regular expressions with counters in some practical applications and we evaluate our algorithm on in synthetic and real benchmark tests based on the expressions we found in practice. Our benchmarks indicate that the new algorithm never performs worse than the state-of-the art but is up to a million times faster when the data is large.

3.15 Logical Definitions of Mildly Context-Sensitive Grammar Formalisms

Uwe Mönnich (Universität Tübingen, DE)

Joint work of Mönnich, Uwe
Main reference U. Mönnich, "A Logical Characterization of Extended TAGs," in Proc. of the 11th Int'l Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+11), pp. 37–45, Paris, 2012.
URL http://alpage.inria.fr/tagplus11/lib/exe/fetch.php?media=tag_1105.pdf

The development of model-theoretic syntax in terms of a logical specification language that is both expressive and manageable presents problems for mono-level approaches. All these approaches suffer from a lack of expressive power in that the family of regular tree languages properly includes all other language families that are captured by the logical formalisms that have been considered in model-theoretic syntax. It is due to this lack of expressive power that grammatical phenomena like cross-serial dependencies in languages like Swiss German or Bambara are beyond the reach of the kind of logical apparatus currently applied to natural language syntax. The talk offers a solution to these problems by integrating a formally unified notion of grammar morphism into the framework of model-theoretic syntax. The approach we present follows Courcelle's extension of Rabin's method of model-theoretic interpretation. This extended version of the classical method of semantic interpretation serves to carve out the exact position of recent linguistic theories like minimalist syntax and (multicomponent) tree adjoining grammar within the family of mildly context-sensitive languages. In the process of determining this position we rely on the linguistically significant affinities between multiple regular tree grammars and simple context-free tree grammars on the one hand and minimalist syntax and tree adjoining grammars on the other. It turns out that the tree languages which are the output of finite-copying top-down tree transducers applied to regular tree languages are exactly the output tree languages of logical tree transducers which are direction preserving in the sense that edges in the output trees correspond to directed paths in the input trees. A similar result holds of (monadic) simple tree transducers which correspond to logical tree transducers which are either direction preserving or inverse direction preserving. Analyzing these results from the perspective of grammar theory leads to the overall conclusion that the formal counterparts of contemporary models of natural language syntax can be characterized in terms of grammar morphisms in the sense of this talk.

3.16 XML Stream Processing Based on Tree Transducer Composition

Keisuke Nakano (The University of Electro-Communications – Tokyo, JP)

Joint work of Nakano, Keisuke; Shin-Cheng Mu
Main reference K. Nakano, S.-C. Mu, "A Pushdown Machine for Recursive XML Processing," in Proc. of the 4th ASIAN Symp. on Programming Languages and Systems (APLAS'06), LNCS, Vol. 4279, pp. 340–356, Springer, 2006.
URL http://dx.doi.org/10.1007/11924661_21

I gave a talk about an application of tree transducer composition, that is, derivation of XML stream processors from XML tree manipulation programs. There are two styles of XML transformation programs: one is tree manipulation (like DOM programming) and another is stream processing (like SAX programming). Tree manipulation style is much easier to write a program than stream processing style. However, a tree manipulation program is less efficient in both time and space than stream processing because it cannot start to output the result before building the whole tree structure in memory. In this talk, I presented a solution to obtain both advantages of two programming styles by deriving XML stream processors from tree manipulation programs. The derivation is based on the theory of tree transducer composition. I developed new composition laws because known composition laws do not work for this particular problem. This work has been presented in APLAS 2004, APLAS 2006, and a technical part of these results is based on my result published in Journal of Theory of Computing Systems in 2009.

3.17 Panic Automata. Yet Another Automata in Quest of Decidability of Mathematical Theories

Damian Niwinski (University of Warsaw, PL)

Joint work of Knapik, Teodor; Niwinski, Damian; Urzyczyn, Pawel; Walukiewicz, Igor
Main reference T.Knapik, D.Niwinski, P.Urzyczyn, I.Walukiewicz, "Unsafe Grammars and Panic Automata," in Proc. of the 32nd Int'l Colloquium on Automata, Languages and Programming (ICALP'05), LNCS, Vol. 3580, pp. 1450–1461, Springer 2005.
URL http://dx.doi.org/10.1007/11523468_117

What features make the theory of a tree – say, the theory in monadic second-order logic (MSO) – decidable? The Rabin 1969 breakthrough result, establishing decidability of the MSO theory of the full binary tree, gave rise to analogous results for various shapes of trees. One way to classify (in general, labeled) trees is to view them as terms over a first-order signature, generated by recursive schemes (sometimes called tree grammars), using typed non-terminals of any order. In this setting, Rabin established the result for level 0, i.e., regular trees, and Courcelle extended it to level 1, i.e., algebraic trees. Knapik, Niwinski, and Urzyczyn [4] extended it further to trees generated by grammars of any level n, but imposing a technical restriction on the use of parameters, called safety. They a posteriori justified the relevant hierarchy of trees, by characterizing it in terms of higher-order pushdown automata. Recall that these automata, introduced by Maslov in 1974, use stacks of stacks of stacks...; the higher-level push operation duplicates the topmost stack of the appropriate level, and the pop operation pops such a stack. An elegant machine/grammar independent characterization of the hierarchy was later given Caucal [2]. In 2005, the KNU together with I.Walukiewicz [5] and independently Aehlig, de Miranda and Ong [1], showed decidability of the MSO theory of trees generated by grammars of level 2 without any restriction. The proof of the former group was based on an enhancement of the second-order pushdown automaton by a cascade popping operation (invented by Pawel Urzyczyn), called panic. Roughly speaking, this operation, guided by the topmost symbol on the stack, reconstructs the stack on which this symbol was put originally, disregarding its future duplications. Further, Luke Ong [6] established (2006) decidability of the MSO theories of trees generated by unrestricted grammars of any level, and extended (in 2008, together with Hague, Murawski, and Serre [3]) the automata-theoretic characterization, introducing collapsible pushdown automata, which generalize panic automata to all levels. It was only in 2011–2012, when Pawel Parys [7, 8] eventually separated the two concepts, by showing that the panic/collapse operation is indeed needed to capture the expressive power of higher-order grammars, or in other words, that the safety is indeed a restriction.

References

1 Aehlig, K., de Miranda J.G., and Ong, L., The monadic second order theory of trees given by arbitrary level-two recursion schemes is decidable. In: Proc. TLCA '05, Springer LNCS 3461 (2005), 39-54.

2 Caucal, D., On infinite terms having a decidable monadic second-order theory. MFCS 2002, 65–176.

3 Hague, M., Murawski, A.S., Ong, C.-H. L., Serre O., Collapsible Pushdown Automata and Recursion Schemes. LICS 2008, 452-461.

4 Knapik, T., Niwiński, D., Urzyczyn, P., Higher-order pushdown trees are easy. FoSSaCS 2002, 205–222.

5 Knapik, T., Niwiński, D., Urzyczyn, P., Walukiewicz, I., Unsafe Grammars and Panic Automata. ICALP 2005, 1450-1461.

6 Ong, C.-H. L., On Model-Checking Trees Generated by Higher-Order Recursion Schemes. LICS 2006, 81-90.

7 Parys, P., Collapse Operation Increases Expressive Power of Deterministic Higher Order Pushdown Automata. STACS 2011, 603-614.

8 Parys, P., On the Significance of the Collapse Operation. LICS 2012, 521-530.

3.18 Recursion Schemes and Pattern Matching

Chih-Hao Luke Ong (University of Oxford, GB)

Joint work of Ong, Chih-Hao Luke; Steven Ramsay
Main reference C.-H. Luke Ong, S. Ramsay, "Verifying Higher-Order Functional Programs with Pattern-Matching Algebraic Data Types," in Proc. of the 38th Annual ACM SIGPLAN-SIGACT Symp. on Principles of Programming Languages (POPL'11), pp. 587–598, ACM, 2011.
URL http://dx.doi.org/10.1145/1926385.1926453

Higher-order model checking is the model checking of infinite trees generated by higher-order recursion schemes (HORS) and related models of computation. We introduced the problem of deciding monadic second-order theories of trees generated by HORS, and discussed recent decidability proofs by Ong and others. Motivated by the standard functional programming idiom of definition by pattern matching, we introduced pattern matching recursion schemes (PMRS) and their safety verification problem. We sketched a semi-complete method that first builds an abstraction of the (undecidable) verification problem in the form of weak PMRS; a solution is then obtained by successive refinement using a CEGAR loop. This is based on a POPL 2011 paper by Luke Ong and Steven Ramsay.

3.19 Descriptional Complexity of Input-Driven Pushdown Automata

Kai T. Salomaa (Queen's University – Kingston, CA)

Joint work of Okhotin, Alexander; Salomaa, Kai
Main reference A. Okhotin, X. Piao, K. Salomaa, "Descriptional Complexity of Input-Driven Pushdown Automata," in H. Bordihn, M. Kutrib, B. Truthe, (Eds.), Languages Alive, LNCS, Vol. 7300, pp. 186–206, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-31644-9_13

Every nondeterministic input-driven pushdown automaton (IDPDA) has an equivalent deterministic IDPDA and the IDPDA languages retain many of the desirable closure properties of regular languages. The IDPDA model is known in the literature also as a visibly pushdown automaton and is mathematically equivalent to the nested word automaton.

It is known that a deterministic automaton equivalent to a nondeterministic IDPDA of size n may need size $2^{\Omega(n^2)}$ (Alur, Madhusudan, J.ACM 2009). This talk surveys recent work on the descriptional complexity of converting a nondeterministic IDPDA to an unambiguous one and of determinizing an unambiguous IDPDA. Also we survey the descriptional complexity of the Boolean operations, concatenation and Kleene star on IDPDA languages.

References

1 Alexander Okhotin, Xiaoxue Piao, Kai Salomaa: Descriptional Complexity of Input-Driven Pushdown Automata. In: Languages Alive (H. Bordihn, M. Kutrib, B. Truthe (Eds.)), Springer Lecture Notes in Computer Science 2012, pp. 186-206 http://dx.doi.org/10.1007/978-3-642-31644-9_13

3.20 Visibly Pushdown Transducers

Frederic Servais (Hasselt University – Diepenbeek, BE)

Joint work of Servais, Frédéric; Reynier, Pierre-Alain; Gauwin, Olivier; Filiot, Emmanuel; Talbot, Jean-Marc; Raskin, Jean-François
Main reference F. Servais, "Visibly Pushdown Transducers," PhD Thesis, Université Libre de Bruxelles, 2011.
URL http://theses.ulb.ac.be/ETD-db/collection/available/ULBetd-09292011-142239/unrestricted/main.pdf

The present work proposes visibly pushdown transducers (VPTs) for defining transformations of documents with a nesting structure. We show that this subclass of pushdown transducers enjoy good properties. Notably, we show that functionality is decidable in PTime and k-valuedness in co-NPTime. While this class is not closed under composition and its type checking problem against visibly pushdown automata is undecidable, we identify a subclass, the well-nested VPTs, closed under composition and with a decidable type checking problem. Furthermore, we show that the class of VPTs is closed under look-ahead, and that the deterministic VPTs with look-ahead characterize the functional VPTs transductions. Finally, we investigate the resources necessary to perform transformations defined by VPTs. We devise a memory efficient algorithm. Then we show that it is decidable whether a VPT transduction can be performed with a memory that depends on the level of nesting of the input document but not on its length.

3.21 FAST: Functional Abstraction of Symbolic Transductions

Margus Veanes (Microsoft – Redmond, US)

Joint work of Veanes, Margus; D'Antoni, Loris
URL http://www.rise4fun.com/Fast/tutorial/guide

We introduce a tree manipulation language and tool, called FAST, which supports trees over infinite alphabets. The core of FAST is based on a combination of state-of-the-art satisfiability modulo theories solving techniques and tree automata and tree transducer algorithms, enabling it to model programs whose input and output can range over any decidable theory. Overall, we strike a balance between expressiveness and precise analysis that works for a large class of tree-manipulating programs.

3.22 Streaming Tree Transducers

Loris d'Antoni (University of Pennsylvania, US)

Joint work of D'Antoni, Loris; Alur, Rajeev
Main reference R. Alur, L. D'Antoni, "Streaming Tree Transducers," in Proc. of the 39th Int'l Colloquium on Automata, Languages, and Programming (ICALP'12) – Part II, LNCS, Vol. 7392, pp. 42–53, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-31585-5_8
URL http://www.cis.upenn.edu/~lorisdan/papers/stt12short.pdf

We introduce Streaming Tree Transducers as an analyzable, executable, and expressive model for transforming strings, unranked and ranked ordered trees, and forests. Given a linear encoding of the input tree, the transducer makes a single left-to-right pass through the input, and computes the output using a finite-state control, a visibly pushdown stack, and a finite number of variables that can store output chunks that can be combined using the operations of string-concatenation and tree-insertion. We prove that the expressiveness of the model coincides with transductions definable using monadic second-order logic (MSO).We establish complexity upper bounds of ExpTime for type-checking and NExpTime for checking functional equivalence for our model. We consider variations of the basic model when inputs/outputs are restricted tostrings and ranked trees, and in particular, present the model of bottom–up ranked-tree transducers, which is the first known MSO-equivalent transducer model that processes trees in a bottom–up manner.

References

1 Rajeev Alur and Loris D'Antoni. *Streaming Tree Transducers.* ICALP (2), 2012, 42-53.

Participants

- Henrik Björklund
University of Umeå, SE
- Johanna Björklund
University of Umeå, SE
- Adrien Boiret
ENS – Paris, FR
- Bruno Courcelle
University of Bordeaux, FR
- Loris d'Antoni
University of Pennsylvania, US
- Frank Drewes
University of Umeå, SE
- Emmanuel Filiot
Université Paris 12, FR
- Zoltan Fülöp
University of Szeged, HU
- Olivier Gauwin
University of Bordeaux, FR
- Daniel Gildea
University of Rochester, US
- Kazuhiro Inaba
Google Japan, JP
- Florent Jacquemard
IRCAM & INRIA – Paris, FR
- Jan Janousek
Czech Technical University, CZ
- Naoki Kobayashi
University of Tokyo, JP
- Marco Kuhlmann
Uppsala University, SE
- Pavel Labath
University of Bratislava, SK
- Aurélien Lemay
University of Lille III, FR
- Sebastian Maneth
NICTA & University of New South Wales, Sydney, AU
- Wim Martens
Universität Bayreuth, DE
- Uwe Mönnich
Universität Tübingen, DE
- Keisuke Nakano
The University of Electro-Communications – Tokyo, JP
- Joachim Niehren
INRIA Nord Europe – Lille, FR
- Damian Niwinski
University of Warsaw, PL
- Chih-Hao Luke Ong
University of Oxford, GB
- Pierre-Alain Reynier
Université de Provence, FR
- Kai T. Salomaa
Queen's Univ. – Kingston, CA
- Helmut Seidl
TU München, DE
- Frédéric Servais
Hasselt Univ. – Diepenbeek, BE
- Jean-Marc Talbot
Aix-Marseille University, FR
- Sophie Tison
Université de Lille I, FR
- Jan Van den Bussche
Hasselt University, BE
- Margus Veanes
Microsoft – Redmond, US
- Heiko Vogler
TU Dresden, DE

Report from Dagstuhl Seminar 13201

Information Visualization – Towards Multivariate Network Visualization

Edited by

Andreas Kerren[1], Helen C. Purchase[2], and Matthew O. Ward[3]

1 **Linnaeus University – Växjö, SE,** kerren@acm.org
2 **University of Glasgow, GB,** helen.purchase@glasgow.ac.uk
3 **Worcester Polytechnic Institute, US,** matt@cs.wpi.edu

Abstract

Information Visualization (InfoVis) focuses on the use of visualization techniques to help people understand and analyze large and complex data sets. The aim of this third Dagstuhl Seminar on Information Visualization was to bring together theoreticians and practitioners from Information Visualization, HCI, and Graph Drawing with a special focus on multivariate network visualization, i.e., on graphs where the nodes and/or edges have additional (multidimensional) attributes. To support discussions related to the visualization of real world data, researchers from selected application areas, especially bioinformatics, social sciences, and software engineering, were also invited. During the seminar, working groups on six different topics were formed and enabled a critical reflection on ongoing research efforts, the state of the field in multivariate network visualization, and key research challenges today. This report documents the program and the outcomes of Dagstuhl Seminar 13201 "Information Visualization – Towards Multivariate Network Visualization".

Seminar 12.–17. May, 2013 – www.dagstuhl.de/13201
1998 ACM Subject Classification E.1 [Data Structures]: Graphs and networks, 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 Information visualization, visualization, network visualization, graph drawing, visual analytics, network analysis, interaction
Digital Object Identifier 10.4230/DagRep.3.5.19
Edited in cooperation with Björn Zimmer (Linnaeus University – Växjö, SE)

1 Executive Summary

Andreas Kerren
Helen C. Purchase
Matthew O. Ward

Introduction

Information Visualization (InfoVis) is a research area that focuses on the use of visualization techniques to help people understand and analyze data. While related fields such as Scientific Visualization involve the presentation of data that has some physical or geometric correspondence, Information Visualization centers on abstract information without such correspondences, i.e., it is not possible to map this information into the physical world in most cases. Examples of such abstract data are symbolic, tabular, networked, hierarchical, or textual information sources.

The first two Dagstuhl Seminars on Information Visualization aimed to cover more general aspects of our field, such as interaction, evaluation, data wrangling, and collaboration, or focused on higher level topics, for instance, the value of InfoVis or the importance of aesthetics. Besides the Dagstuhl reports that are typically published directly after a seminar [1, 2, 4, 5], there were also follow-up publications for both seminars. The participants of Seminar #07221 wrote book chapters which have been consolidated into a Springer book [7]; the organizers of the same seminar published a workshop report in the Information Visualization journal [6]. For the second Seminar #10241, a special issue in the same journal was published [3].

The goal of this third Dagstuhl Seminar on Information Visualization was to bring together theoreticians and practitioners from Information Visualization, HCI, and Graph Drawing with a special **focus on multivariate network visualization**, i.e., on graphs where the nodes and/or edges have additional (multidimensional) attributes. The integration of multivariate data into complex networks and their visual analysis is one of the big challenges not only in visualization, but also in many application areas. Thus, in order to support discussions related to the visualization of real world data, we also invited researchers from selected application areas, especially bioinformatics, social sciences, and software engineering. The unique "Dagstuhl climate" ensured an open and undisturbed atmosphere to discuss the state-of-the-art, new directions, and open challenges of multivariate network visualization.

Seminar Topics

The following themes were discussed during the seminar. The seminar allowed attendees to critically reflect on current research efforts, the state of field, and key research challenges today. Participants also were encouraged to demonstrate their system prototypes and tools relevant to the seminar topics. In consequence, topics emerged in the seminar week and were the focus of deeper discussions too.

- **Focus on biochemistry/bioinformatics:** In the life sciences, huge data sets are generated by high-throughput experimental techniques. Consequently, biologists use computational methods to support data analysis. The information in many experimental data sets can be either represented as networks or interpreted in the context of various networks. How can our current techniques help to analyze primary and secondary data in the context of such networks, and how can different network types be combined?
- **Focus on social science:** Graph drawing techniques have been used for several years for the visualization and analysis of social networks, but other social science fields (e.g., geography, politics, cartography, and economics) also make use of data visualization. How can (or do) our network visualizations support these domains?
- **Focus on software engineering:** In the application domain of software engineering, various graphs and data attached to graphs (e.g., software metrics) play a dominant role in the static and dynamic analysis of programs. Which of these problems are conceptually similar to graph-related problems in biology or social sciences and how can multivariate network visualization support specific tasks, such as software architecture recovery?
- **Approaches and methods:** There already exist a number of technical approaches, algorithms, and methods to interactively visualize multivariate networks. Which ones are suitable for solving specific tasks in our applications areas? What is their potential? What are their limitations? By identifying the range of approaches that do exist, can we see the potential for new, innovative visualization ideas?

- **Challenges in visualizing multivariate networks:** Multivariate networks are large and complex and their complexity will increase in the future. Thus, not all problems can be solved in the short term. What are the current challenges?
- **Time-dependent/dynamic networks:** Many networks that are considered in practice change over time with respect to their topology and/or their attributes. How can we best visualize networks and attributes that change over time?
- **Interaction:** How can we best support the navigation, exploration and modification of multivariate networks?
- **Multiple networks at different scales:** How can we integrate, combine, compare more than one multivariate network at different scales? In this context, the term of so-called multi-modal networks is often used in literature. What does this term mean exactly? Can we visualize a range of different information types concurrently?
- **Tasks:** What range of tasks can multivariate network visualization support? Are there general tasks for all application domains?
- **Novel metaphors:** What type of visualization metaphors should we use beyond node-link diagrams? What would be the benefit in doing so?

Outcomes

The organizers and participants decided to write a book on multivariate network visualization to be published as LNCS issue by Springer. The possibility of publishing this Springer book was confirmed by the Editor-in-Chief of LNCS already before the start of the seminar. Working groups have been invited to submit a book chapter building on their discussions and findings, and writing is underway. The final chapters are to be submitted by November 3, 2013, with a planned publication date of Spring 2014. A preliminary book structure was presented at the end of the seminar:

1. Introduction
   a. Definition of multivariate networks, typical representations
2. Domain Application Data Characteristics in Context of Multivariate Networks
   a. Biology
   b. Social Sciences
   c. Software Engineering
3. Tasks
4. Interaction
5. Metaphors (Visual Mappings beyond Node-Link)
6. Multiple and Multi-Domain Networks
7. Temporal Networks
8. Scalability
9. Summary/Conclusion

The Dagstuhl team performed an evaluation at the end of the seminar week. The results of this survey (scientific quality, inspiration to new ideas/projects/research/papers, insights from neighboring fields, ...) were throughout very good to excellent. Only a few single improvements were proposed by participants, for example, more junior researchers should be invited to come into contact with world-class researchers. And more domain experts should be invited to be spread out across the breakout groups. Another issue was that the time available for group work should be extended in future seminars.

References

1 Andreas Kerren, Catherine Plaisant, and John T. Stasko. 10241 Abstracts Collection: Information Visualization. In Andreas Kerren, Catherine Plaisant, and John T. Stasko, editors, *Information Visualization*, number 10241 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2010. Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, Germany.

2 Andreas Kerren, Catherine Plaisant, and John T. Stasko. 10241 Executive Summary: Information Visualization. In Andreas Kerren, Catherine Plaisant, and John T. Stasko, editors, *Information Visualization*, number 10241 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2010. Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, Germany.

3 Andreas Kerren, Catherine Plaisant, and John T. Stasko. Information Visualization: State of the Field and New Research Directions. *Information Visualization*, 10(4):269–270, 2011.

4 Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North. 07221 Abstracts Collection: Information Visualization – Human-Centered Issues in Visual Representation, Interaction, and Evaluation. In Jean-Daniel Fekete, Andreas Kerren, Chris North, and John T. Stasko, editors, *Information Visualization – Human-Centered Issues in Visual Representation, Interaction, and Evaluation*, number 07221 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2007. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

5 Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North. 07221 Executive Summary: Information Visualization – Human-Centered Issues in Visual Representation, Interaction, and Evaluation. In Jean-Daniel Fekete, Andreas Kerren, Chris North, and John T. Stasko, editors, *Information Visualization – Human-Centered Issues in Visual Representation, Interaction, and Evaluation*, number 07221 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2007. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

6 Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North. Workshop Report: Information Visualization – Human-Centered Issues in Visual Representation, Interaction, and Evaluation. *Information Visualization*, 6(3):189–196, 2007.

7 Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North, editors. *Information Visualization: Human-Centered Issues and Perspectives*, volume 4950 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2008.

2 Table of Contents

Working Groups

3 Seminar Program and Activities

Participation and Program

41 people from 9 countries participated in this seminar. Most attendees came from the US, United Kingdom and Germany; others came from France, The Netherlands, and other European countries as shown in Figure 1. At least two domain experts from the application domains (bioinformatics, social sciences, software engineering) participated in the seminar.

The program aimed to generate lively discussions. Before the seminar started, the organizers asked the domain experts to prepare talks (45 minutes) that highlight the characteristics of their domain-specific data sets and the tasks that analysts perform on the data. At the same time, all participants were called to volunteer for three survey talks (45 minutes) which show the state-of-the-art of visualization and interaction techniques applied to the data considered in each domain; in addition three talks of the same length were invited to present some of the technical challenges of multivariate network visualization within each domain. Individual presentations on new techniques were also welcome.

Thus, the first half of the seminar week was mainly dedicated to the presentations of the invited talks followed by discussions. In the second half, primarily breakout groups were formed to deal with specific topics (briefly illuminated in the next subsection) including group reports to the whole audience. On Thursday, the organizers alloted two slots for scientific presentations on techniques and tool demos. Table 1 provides an overview of the final seminar schedule.

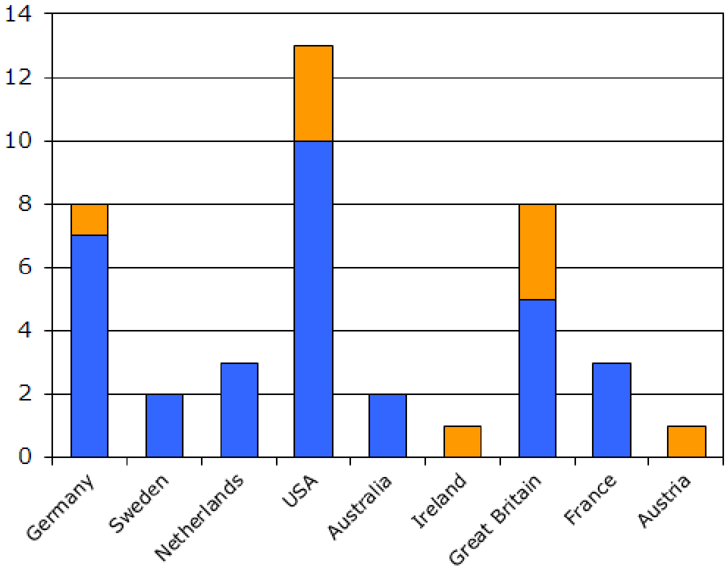


Figure 1 Attendee Statistics of Seminar #13201. Blue colored bars represent male and orange colored bars female participants.

Table 1 Final structure of the seminar. The main discussion topic of Monday was network visualization with a focus on bioinformatics. Tuesday was mainly focused on social network visualization, whereas Wednesday morning was used to discuss visualizations in context of networks that occur in software engineering. Thursday and Friday were mainly focused on general techniques for the visualization of multivariate networks as well as on group work and group reporting.

Monday	Tuesday	Wednesday	Thursday	Friday
Welcome Self Presentations	Characteristics-Talks (social) Presentations (social, 1 talk)	Characteristics/ Survey-Talk (SE) Challenges-Talk (SE)	Presentations (general, 3 talks)	Synergy Session & Sum Up
Self Presentations Characteristics-Talks (bio)	Challenges-Talk (social) Presentations (social, 2 talks)	Breakout Groups Group Reporting	Breakout Groups	Book Planning & Closing Remarks
Survey-Talk (bio) Challenges-Talk (bio)	Discussion: Topics for Breakout Groups	Social Event (Völklinger Hütte)	Presentations/ Demos (general, 4 talks)	
Presentations (general, 3 talks)	Breakout Groups		Group Reporting	

Activities

Invited Talks

The titles and presenters of invited talks for each application domain are listed in the following. Abstracts for the individual talks can be found in Sect. 4.

- Multivariate Networks in Biology
  - *Matthew O. Ward and Carsten Görg:* Biological Multivariate Network Visualizations – A Partial Survey [survey]
  - *Oliver Kohlbacher:* Characteristics of Biological Data in the Age of Omics – Part 1 [characteristics]
  - *Falk Schreiber:* Characteristics of Biological Data – Part 2 [characteristics]
  - *Jessie Kennedy, Oliver Kohlbacher, and Falk Schreiber:* Challenges in Visualising Biological Data [challenges]
- Multivariate Networks in Social Sciences
  - *Lothar Krempel:* Social Science Applications of Multivariate Network Visualization [characteristics]
  - *Maura Conway:* Visualising Terrorism and Violent (Online) Political Extremism Data [characteristics]
  - *Michelle X. Zhou:* Top Challenges in Visualizing Multivariate Personal Networks [challenges]
- Multivariate Networks in Software Engineering
  - *Stephan Diehl and Alexandru C. Telea:* Multivariate Graphs in Software Engineering [characteristics, survey, challenges]

Breakout Groups

As already mentioned above, the program included breakout sessions on six specific topics, i.e, six working groups discussed one topic at a time in parallel sessions. The themes were based on topics discussed in the original seminar proposal as well as topics that emerged in the first session on Tuesday afternoon. The detailed working group reports are presented in Sect. 5. In the following, we list the different groups:

1. Temporal Multivariate Networks
2. Interaction for Multivariate Networks
3. Multiple and Multi-domain Networks
4. Scalability of Multivariate Graph Visualization
5. Tasks for Multivariate Network Analysis
6. Novel Visual Metaphors for Multivariate Networks

Scientific Talks and Demos

In addition, a number of speakers gave a scientific talk and/or a tool demo on a theme related to the research questions of the seminar. In sum, 14 talks/demos were given during the seminar (cf. Sect. 4 for details):

- *James Abello:* Multivariate Time-Variant Graphs and Simplicial Complexes: Work in Progress
- *Daniel Archambault:* Effective Visualization of Information Cascades
- *Katy Börner:* The Information Visualization MOOC
- *Tim Dwyer:* CodeMap: Visualizing Software Dependencies in Microsoft Visual Studio
- *Peter Eades:* On the Faithfulness of Graph Visualisations
- *Niklas Elmqvist:* Multi-Phase/Variate/Modal Graphs: Visualization, Interaction, and Evaluation
- *Benjamin David Hennig:* The Complexity of Space
- *Christophe Hurter:* Scalable Multivariate Data Exploration Tools
- *Stephen G. Kobourov:* Maps of Computer Science
- *Kwan-Liu Ma:* Visualization for Studying Social Networks
- *Silvia Miksch:* A Matter of Time: Visual Analytics of Dynamic Social Networks
- *Martin Nöllenburg:* Column-based Graph Layout for Argument Maps
- *John T. Stasko:* Multivariate Network Data & Attribute-based Layout: Two Examples
- *Kai Xu:* GraphScape: Integrated Multivariate Network Visualization

The content of these talks, given for all seminar attendees, raised further key issues and helped the groups to discuss their individual theme from various perspectives.

4 Overview of Talks

4.1 Multivariate Time-Variant Graphs and Simplicial Complexes: Work in Progress

James Abello (Rutgers University – Piscataway, US)

One of the ultimate goals of time-variant data analysis is to synthesize the structure, behavior, and evolution, of the encoded information. Folklore evolution and information flow in communication networks constitute useful data sources to identify some fundamental time-variant data facets. On the other hand, simplicial complexes are useful mathematical tools that aid in the formulation and identification of some of the essential computational questions associated with time-variant graph data. In this case, contractions to planar graphs offer visual appealing representations of the data space that can be used as the main mechanisms for data interaction. These slides offer quick entry points to Computational Folkloristics, Discrepancy in Communication Networks, and Planar Graph Contractability.

4.2 Effective Visualization of Information Cascades

Daniel Archambault (Swansea University, GB)

Dynamic graph attributes appear in many applications, including social media analysis. An attribute is a value associated with a node or edge of the graph and a dynamic attribute is one that changes its value over time. In graphs used in social media analysis, nodes may receive and transmit dynamic attributes to their neighbors, and long chains of these transmissions are known as cascades. Typical methods for visualizing cascades include animations and small multiples representations where nodes change color with the changing attribute value. We present the results of a formal user study that tests the effectiveness of dynamic attribute visualization on graphs. We test the task of locating nodes which amplify the propagation of a cascade and factors such as animation and small multiples, and force-directed and hierarchical layouts. Overall, we found that small multiples was significantly faster than animation with no significant difference in terms of error rate. Participants generally preferred animation to small multiples and a hierarchical layout to a force-directed layouts. Considering each presentation method separately, when comparing force-directed layouts to hierarchical layouts, hierarchical layouts were found to be significantly faster for both presentation methods and significantly more accurate for animation. Thus, for our task, this experiment supports the use of a small multiples interfaces with hierarchically drawn graphs for the visualization of dynamic attributes.

4.3 The Information Visualization MOOC

Katy Börner (Indiana University – Bloomington, US)

The talk discusses the structure and design of the Information Visualization MOOC taught in Spring 2013 and students from 93 different countries attended. Among other topics, the course covers:

- Data analysis algorithms that enable extraction of patterns and trends in data.
- Major temporal, geospatial, topical, and network visualization techniques.
- Discussions of systems that drive research and development.

Everybody who registers at http://ivmooc.cns.iu.edu gains free access to the Scholarly Database (26 million paper, patent, and grant records) and the Sci2 Tool (100+ algorithms and tools). First 'learning analytics' visualizations of student activity and collaboration networks are presented as well.

4.4 Visualizing Terrorism and Violent (Online) Political Extremism Data

Maura Conway (Dublin City University, IE)

This presentation explored the ways in which data on terrorism and violent (online) political extremism has been visualized to-date and ways in which the InfoVis community could contribute in this research area going forward. There are three major reasons for visualizing terrorism and violent political extremism data:

1. To answer specific questions.
2. To explore the data and thereby potentially identify new avenues for research and analysis.
3. To communicate research and results to other researchers, policy makers, security professionals, and others.

Some well-known visualizations in this area include those of specific terrorist networks, including the 9/11 attackers social network(s) (Krebs 2002); time-lapse visualizations of the event data contained in the Global Terrorism Database (GTD); and mapping of various aspects of the Northern Ireland conflict by the CAIN project. The presentation focused, in particular, on the 'new' data available to researchers into terrorism and violent political extremism resulting from the latter migrating some of their activity to the Internet, including social networking sites, which has not been extensively visualized to-date. Suggestions for ways in which the InfoVis community could productively collaborate with scholars in the field of terrorism and violent (online) political extremism were then identified, including mapping terrorist attacks and incidences of violent political extremism; visualizing online VPE content distribution networks; visualizing VPE data collected from across online social networking sites, and similar.

4.5 Multivariate Graphs in Software Engineering

Stephan Diehl (Universität Trier, DE) and Alexandru C. Telea (University of Groningen, NL)

In this presentation we first talk about the importance of visualization in software engineering in general. Next, we give examples of the typical kinds of data used in software engineering in terms of items, attributes, and relations. Since all these can change over time, we conclude that dynamic multivariate compound graphs provide a good common model. Next, we present many examples of visualization tools that show multivariate data and dynamic graphs related to the structure, behavior and evolution of software. For each tool, we give a typical task to be solved with the tool. We conclude by suggesting some challenges for future research in the area.

4.6 CodeMap: Visualizing Software Dependencies in Microsoft Visual Studio

Tim Dwyer (Monash University Melbourne, AU)

I worked for four years with the Visual Studio product group at Microsoft to develop a software dependency visualization tool. In that time we conducted many user studies to try to better understand how graph visualization can help developers with their most difficult tasks. Based on the feedback from these studies we iterated a tool that attempts to support a developer's working memory with a visual scratchpad that co-exists with their established, code-centric workflow. In this talk I will demonstrate the tool we have developed: CodeMap. I hope to share some of the insights we gained from our studies, which sometimes ran counter to own expectations and some of the tacit assumptions of the visualization fraternity.

4.7 On the Faithfulness of Graph Visualisations

Peter Eades (The University of Sydney, AU)

We argue that the classical readability criteria for visualizing graphs, though necessary, are not sufficient for effective graph visualization. We introduce another kind of criterion, generically called "faithfulness", that we believe is necessary in addition to readability. Intuitively, a graph drawing algorithm is "faithful" if it maps different graphs to distinct drawings. In other words, a faithful graph drawing algorithm never maps distinct graphs to the same drawing. This concept is extended to a task-oriented model of visualization.

4.8 Multi-Phase/Variate/Modal Graphs: Visualization, Interaction, and Evaluation

Niklas Elmqvist (Purdue University, US)

In this talk, I will review our recent work on visualizing, navigating in, and interacting with various types of complex graphs: those that (1) evolve over time, (2) contain multivariate data, and (3) involve different types of nodes and links. Specific projects include TimeMatrix, GraphDice, Dynamic Insets, the COE Explorer, and Parallel Node-Link Bands. I will also discuss several human subjects evaluations we performed on various graph-related tasks. Taken together, these projects begin to form an outline for the design space of complex (multi-*) graph visualization.

4.9 The Complexity of Space

Benjamin David Hennig (University of Sheffield, GB)

How can we reduce the complexity of flows to a significant visual representation relating to our basic geographic understanding of the world? From a geographic perspective, this relates strongly to theoretical and conceptual questions of space, which can be understood in manifold ways. Transferring such concepts into geographic visualizations, as demonstrated with a novel gridded approach to applying density-equalizing map transformations, may provide a useful new idea to visualizing flows in other ways. This talk outlines the thinking behind these ideas and presents some examples of how this might contribute to new ways of showing geographic flows.

4.10 Scalable Multivariate Data Exploration Tools

Christophe Hurter (ENAC – Toulouse, FR)

Interactive data exploration and manipulation are often hindered by the size of the datasets. This is particularly true for 3D datasets where the problem is exacerbated by occlusion, important adjacencies, and entangled patterns. These complexities make visual interaction via common filtering techniques difficult. In this presentation, I described a set of techniques aimed at performing real-time multi-dimensional data deformation with the intention of helping people to easily select, analyze, and eliminate specific spatial-and data patterns. Our interaction techniques allow animation between view configurations, semantic filtering and view deformation. Any subset of the data can be selected at will at any step along the animation. Selected data can be filtered and deformed in order to remove occlusion and ease complex data selections. I applied our techniques to the following domain areas: 3D medical imaging, and multivariate network. The technique is simple, flexible and interactive with large datasets (up to 50 of millions displayed data points).

4.11 Challenges in Visualising Biological Data

Jessie Kennedy (Edinburgh Napier University, GB), Oliver Kohlbacher (Universität Tübingen, DE), and Falk Schreiber (IPK Gatersleben & MLU Halle, DE)

Past Dagstuhl Seminars already identified some challenges in visualising biological networks, as described in the paper "A graph-drawing perspective to some open problems in molecular biology" by Albrecht et al. (GD '09). In this presentation, we extend this work by other technical challenges (such as the visual analysis of ontologies or the important aspect of uncertainty in context of multivariate networks) as well as personal challenges of visualization researchers (like the need to understand biological problems or communicating with biologists). Open visualization problems are highlighted with the help of many practical tool examples.

4.12 Maps of Computer Science

Stephen G. Kobourov (University of Arizona – Tucson, US)

We describe a practical approach for visual exploration of research papers. Specifically, we use the titles of papers from the DBLP database to create maps of computer science (MOCS). Words and phrases from the paper titles are the cities in the map, and countries are created based on word and phrase similarity, calculated using co-occurrence. With the help of heatmaps, we can visualize the profile of a particular conference or journal over the base map. Similarly, heatmap profiles can be made of individual researchers or groups such as a department. The visualization system also makes it possible to change the data used to generate the base map. For example, a specific journal or conference can be used to generate the base map and then the heatmap overlays can be used to show the evolution of research topics in the field over the years. As before, individual researchers or research groups profiles can be visualized using heatmap overlays but this time over the journal or conference base map. Finally, research papers or abstracts easily generate visual abstracts giving a visual representation of the distribution of topics in the paper. We outline a modular and extensible system for term extraction using natural language processing techniques, and show the applicability of methods of information retrieval to calculation of term similarity and creation of a topic map. The system is available at mocs.cs.arizona.edu.

4.13 Characteristics of Biological Data in the Age of Omics – Part 1

Oliver Kohlbacher (Universität Tübingen, DE)

This talk gives a brief overview of omics data and its relationship to networks. First, I introduce the usual terminology and show the differences between classical data and omics data. Omics data in systems biology either represent a network or is interpreted in the context of a network. Based on these fundamental characteristics, a hierarchy of biological networks can be identified. The various network elements at the different levels in the hierarchy differ not only in their meaning, but also in their scale.

4.14 Social Science Applications of Multivariate Network Visualization

Lothar Krempel (MPI für Gesellschaftsforschung – Köln, DE)

The social network perspective seeks to understand how the embeddedness of actors provides specific opportunities for action (social capital). Network data allow to characterize actors according to their local or global centrality, their status or equivalent roles, to identify clusters, cliques or communities. These metrics can help to understand how social systems work. Additional attributes of the actors or their relations can greatly enhance the analysis.

Visualizations are an indispensable tool to inspect such complex datasets. I present results from scientific cooperations, where we able to gain new insights into social science questions. Data on capital ties among the largest companies—as collected by governmental agencies—can be used to trace a historic process, i.e., how national economies transform under the regime of internationalization and financial liberalization. Letters among scientists in the 18th century can be used to generate historical science maps. Growth rates in car trade and the composition of trade in parts and components vs. assembled cars give detailed insight into the state of an international division of labor. Data on awards to universities per funding area allow for insight into the specialization of German universities.

4.15 Visualization for Studying Social Networks

Kwan-Liu Ma (University of California – Davis, US)

The network datasets produced by studies in social science have several characteristics. The datasets usually contain collections of small networks. They are mostly multivariate and categorical. There are often missing, incomplete data due to the nature of the data collection methods. Studying the resulting data suggest the need of egocentric visualization techniques. In my presentation, I share with you my experience in working with sociologists to develop egocentric visualization techniques for studying their data. I describe the particular datasets I have worked on and the visualization designs we have derived to meet their needs. As the data collection methods are changing with the widespread use of mobile and web applications, we anticipate a rapid growth in data size in the near future. I also discuss a few techniques for simplifying and visualizing large networks.

4.16 A Matter of Time: Visual Analytics of Dynamic Social Networks

Silvia Miksch (TU Wien, AT)

Several techniques have been proposed to examine static network data, but the visual analysis of dynamic network data is an emerging research field with several open questions. The dynamic and multi-relational nature of this data poses the challenge of understanding both its topological structure and how it changes over time. In this talk, I will present the applied research project ViENA (Visual Enterprise Network Analytics) and how a visual analytics approach supports the examination of dynamic networks according to specific user tasks. Finally, I show possible future directions and challenges.

4.17 Column-based Graph Layout for Argument Maps

Martin Nöllenburg (KIT – Karlsruhe Institute of Technology, DE)

An argument map is a diagram that captures the logical structure of the various arguments and statements in a debate. It can be modeled as a directed hierarchical graph, where vertices are represented as variable-size boxes containing components of an argument and edges indicating relationships as support or attack. We present a new algorithm of Betz et al. (GD'12) for column-based argument map layout following the topology-shape-metrics framework. It uses rectangular fixed-width boxes and orthogonal polyline edges with at most 4 bends. The algorithm is implemented in the Argunet tool (www.argunet.org), of which we show a short demo.

4.18 Characteristics of Biological Data – Part 2

Falk Schreiber (IPK Gatersleben & MLU Halle, DE)

Today, it is of great importance to combine omics, image, volume, and network data with mapping, analysis and visualization possibilities. In this talk, I show the problem of multi-domain biological data and the need of an integrated analysis and visualization of such multi-domain biological data.

4.19 Multivariate Network Data & Attribute-based Layout: Two Examples

John T. Stasko (Georgia Institute of Technology, US)

In this talk I present short demos of two systems. The first, dotlink360, visualizes financial and descriptive information about companies in the mobile computing ecosystem. More importantly, it visualizes alliances and partnerships between these companies. Among others, the system presents four network visualizations using attribute-based node positioning. The second demo presents IEEE InfoVis Conference papers. It shows the total number of Google Scholar citations for each, along with internal citations between the papers. Our technique does not draw edges between the papers but instead uses interaction to highlight the references.

4.20 Biological Multivariate Network Visualizations – A Partial Survey

Matthew O. Ward (Worcester Polytechnic Institute, US) and Carsten Görg (University of Colorado, US)

In this talk, we present a variety of examples for multivariate network visualizations of biological data. We first review a number of different encodings used to map multivariate data to network nodes. We then introduce plugins for the Cytoscape network visualization framework, including the Hanalyzer plugin for exploring knowledge networks in the context of experimental data, the MetaMapp plugin for identifying differentially regulated metabolites, and the Cerebral plugin for analyzing multiple experimental conditions on a graph with biological context. We also review stand-alone tools, including Chilibot for identifying relationships between genes and proteins in the biomedical literature, enRoute for context preserving pathway visualization and analysis, and a system for uncertainty-aware visual analysis of biochemical reaction networks. We describe the networks according to their node mappings, edge mappings, and analysis tasks they support.

4.21 GraphScape: Integrated Multivariate Network Visualization

Kai Xu (Middlesex University, GB)

We introduce a method, GraphScape, to visualize multivariate networks, i.e., graphs with multivariate data associated with their nodes. GraphScape adopts a landscape metaphor with network structure displayed on a 2D plane, and the surface height in the third dimension represents node attribute. More than one attribute can be visualized simultaneously by using multiple surfaces. In addition, GraphScape can be easily combined with existing methods to further increase the total number of attributes visualized. One of the major goals of GraphScape is to reveal multivariate graph clustering, which is based on both network structure and node attributes. This is achieved by a new layout algorithm and an innovative way of constructing attribute surfaces, which also allows visual clustering at different scales through interaction.

4.22 Top Challenges in Visualizing Multivariate Personal Networks

Michelle X. Zhou (IBM Almaden Center – San José, US)

Given the abundant digital footprints left by people, we are developing people analytics and engagement technologies to facilitate deeper understanding of people and to optimize individualized engagements (e.g., doctors-patients and marketers-consumers). To help users either gain a better understanding of themselves or others, we use a personal network to represent each individual. In a personal network, each individual is a multivariate node, of which traits are automatically derived from one's linguistic footprints, while each link represents the individual's social contact also automatically extracted and characterized from one's interaction and communication activities with others. Based on the characteristics of data, value, variety, veracity, velocity, volume, and user capabilities, I list a set of challenges in creating, exploring, and improving personal networks.

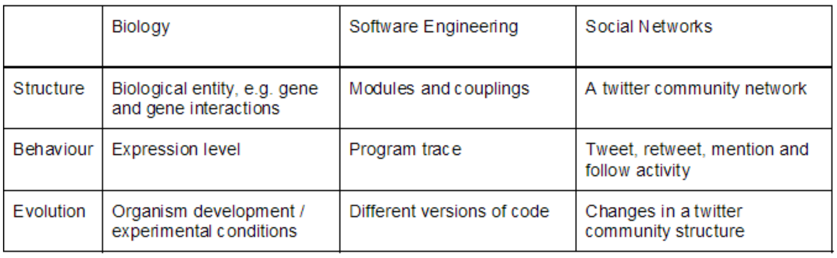
5 Working Groups

This section describes results from each of the six working groups and identifies the attendees contributing to each group. The names of those people who reported for the working groups (group leaders) are underlined.

5.1 WG: Temporal Multivariate Networks

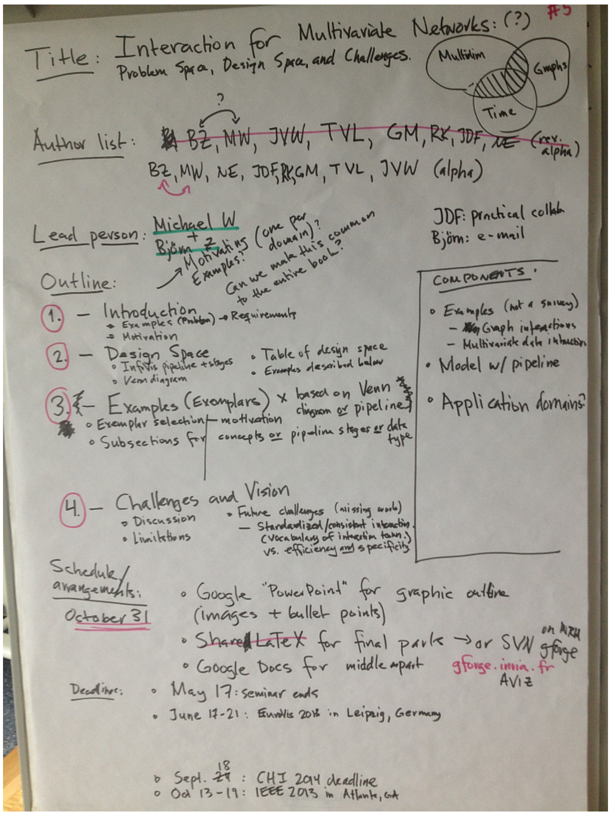
James Abello, Daniel Archambault, Maura Conway, Stephan Diehl, Carsten Görg, Benjamin David Hennig, Jessie Kennedy, Stephen G. Kobourov, Lothar Krempel, Kwan-Liu Ma, Silvia Miksch, and Alexandru C. Telea

The breakout group aimed to characterize the temporal dimension in visualizing multivariate networks exemplified through the study of three application domains that were discussed in the seminar (biology, software engineering and social sciences). As a first step, we classified characteristic domain examples of networks into three different categories (cf. Figure 2): the *structure* describes arrangement and relationships between parts or elements (i.e., the network itself). *Behavior* means the action or reaction of something under given circumstances, and *evolution* is the gradual development of something over time. Based on this underlying model, the group members plan to structure their book chapter as follows: after a presentation of the various types of graphs for temporal networks, we try to specify typical tasks that usually have to be performed with temporal graphs. Then, a domain characterization will be discussed as already mentioned above. Finally, the group will do a survey/classification of existing visualization techniques for temporal multivariate networks followed by highlighting visualization challenges in this context.



	Biology	Software Engineering	Social Networks
Structure	Biological entity, e.g. gene and gene interactions	Modules and couplings	A twitter community network
Behaviour	Expression level	Program trace	Tweet, retweet, mention and follow activity
Evolution	Organism development / experimental conditions	Different versions of code	Changes in a twitter community structure

Figure 2 Domain examples classified according to structure, behavior and evolution.


#5
Multidim
Graphs
Time
?

Title: Interaction for Multivariate Networks: (?)
Problem Space, Design Space, and Challenges.

Author list: ~~BZ, MW, JVW, TVL, GM, RK, JDF, NE (rev. alpha)~~
BZ, MW, NE, JDF, RK, GM, TVL, JVW (alpha)

Lead person: Michael W + Björn Z
Motivations (one per domain)? Examples?
Can we make this common to the entire book?

JDF: practical collab.
Björn: e-mail

Outline:

1. – Introduction
   - Examples (Problem) → Requirements
   - Motivation
2. – Design Space
   - InfoVis pipeline + stages
   - Venn diagram
   - Table of design space
   - Examples described below
3. – Examples (Exemplars) × based on Venn diagram or pipeline
   - Exemplar selection – motivation
   - Subsections for concepts or pipeline stages or data type
4. – Challenges and Vision
   - Discussion
   - Limitations
   - Future challenges (missing work)
     – Standardized/consistent interaction (vocabulary of interaction terms) vs. efficiency and specificity

COMPONENTS:
- Examples (not a survey)
  – Graph interactions
  – Multivariate data interaction
- Model w/ pipeline
- Application domains?

Schedule/arrangements:
October 31
- Google "PowerPoint" for graphic outline (images + bullet points)
- ~~SharedLaTeX~~ for final parts → or SVN on INRIA gforge
- Google Docs for middle part
gforge.inria.fr
AVIZ

Deadlines:
- May 17: seminar ends
- June 17-21: EuroVis 2013 in Leipzig, Germany
- Sept. ~~24~~ 18: CHI 2014 deadline
- Oct 13-19: IEEE 2013 in Atlanta, GA

Figure 3 Possible structure of the corresponding "interaction" chapter.

5.2 WG: Interaction for Multivariate Networks

Niklas Elmqvist, Jean-Daniel Fekete, Robert Kosara, Guy Melançon, Jarke J. van Wijk, Tatiana von Landesberger, Michael Wybrow, and Björn Zimmer

The interaction group started out discussing existing interaction techniques for networks and multivariate data sets to order to get an overview of the field. The group examined in greater detail the problems and requirements that arise during the analysis process of graphs with multivariate attributes. It was decided to categorise interaction techniques by their place in the stages of the standard InfoVis pipeline (cf. Figure 3), i.e., whether they operate at data, visual encoding or view level. The group planned and discussed the overall structure and content for the interaction book chapter and decided how they would continue to work on it after the seminar. It was decided that the chapter would not just take the form of a survey of existing techniques, but rather be a guide to the use of interaction in the visualisation of multivariate networks; describing various popular interaction techniques, discussing application of novel approaches, and exploring the remaining challenges in this space.

5.3 WG: Multiple and Multi-domain Networks

Katy Börner, Hans Hagen, Andreas Kerren, and Falk Schreiber

Multiple networks and multi-domain networks occur in several application fields, where their integration, combination, comparison, and visualization is one of the big challenges. The working group started to analyze the general characteristics of this type of data and identified examples in the three application domains of the seminar: biology, social sciences, and software engineering. Figure 4 shows the outcome of this discussion. Conceptually, there are sets of multivariate networks at two or more levels. Each level may describe a specific scale, and within each level several related multimodal or heterogeneous networks are represented. To take an example from biology: within a level different molecular-biological networks such as metabolism, protein interaction and gene regulation networks may be integrated, across levels the scale could be from a molecular biological network to an evolutionary network. We allow $n : m$ mappings within the same level, but only $1 : n$ mappings across levels that must be consecutive. This leads to a structured data set that is the basis for further visual analyses. At the end of the seminar, the group proposed a set of example visual representations, discussed application domain examples, and presented the structure of the planned book chapter.

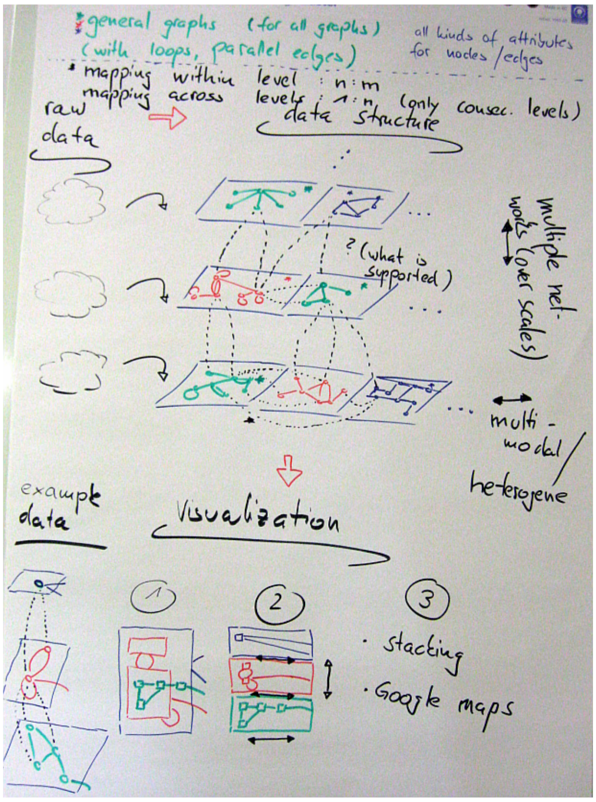

* general graphs (for all graphs)
(with loops, parallel edges)
all kinds of attributes for nodes / edges
* mapping within level : n : m
mapping across levels : 1 : n (only consec. levels)
raw data
data structure
? (what is supported)
multiple networks (over scales)
multi-modal
heterogene
example data
visualization
1
2
3
· stacking
· Google maps

Figure 4 Sketch to understand the fundamental problem of analyzing multiple networks at different scale.

5.4 WG: Scalability of Multivariate Graph Visualization

Tim Dwyer, Danny Holten, Christophe Hurter, T. J. Jankun-Kelly, Martin Nöllenburg, and Kai Xu

The working group for Scalability worked under the mantra "Draw what you need when you need it". As a starting off point, we assumed that Shneiderman's visual seeking mantra needs a caveat: choose an overview that is appropriate to the task—i.e. rarely do users want to see the whole universe. Large (especially scale-free graphs) inevitably turn into hairballs—layout can help: but it happens sooner or later even to the best layout methods. Matrix views do not have the problem of edge crossings, but they have a limited ability to convey transitive structure, also the amount of display area they require grows in the square of the number of nodes. Thus, we end up with displays of hairballs or white noise. Multivariate structures just add further complication. Thus, to understand the issues of complexity in scalable multivariate graphs, we focused our discussion on five dimensions: the display resolution (2- vs. 3-dimensions), the limitations of visual acuity on dense graph perception, the related limitations of cognitive resources in understanding complex graphs, software architectural limitations on the processing very large graphs, and how other visual dimensions (such as color or over plotting) can affect our understanding. These dimensions will be used to build up suggestions for grounded approaches for effectively using large, multivariate graphs.

5.5 WG: Tasks for Multivariate Network Analysis

Peter Eades, Helen Gibson, Daniel A. Keim, A. Johannes Pretorius, Helen C. Purchase, and John T. Stasko

This workgroup set out to characterize the tasks associated with multivariate network analysis. A *task* was considered to be an activity that a user wishes to accomplish by means of interaction with a visual representation of a multivariate network.

All workgroup participants agreed that a distinction could be made between *low-level* tasks and *high-level* tasks. The group felt that this characterization is independent of the application domain. Low-level tasks are atomic interactive analysis activities. Examples include retrieving an attribute value of a node, or determining if nodes are related to each other. High-level tasks cater more specifically for the relational and multivariate nature of multivariate networks and are composed of low-level tasks. Examples include analyzing the overall structure of the network, or finding clusters of nodes that have specific attribute values.

Other issues discussed included interaction as a means of performing tasks (a topic we felt might best be left to the interaction group) and network comparisons (a topic that might best be included in the work of the temporal or multiple network groups).

Having thus defined the scope of our "task" chapter, we can draw on the existing literature that defines tasks for (non-multivariate) graphs and other visualization methods. This literature also often distinguishes between low- and high-level tasks. From our discussions it emerged that there may be another level of tasks somewhere between low- and high-level. However, this issue was not resolved as it proved to be very difficult to map to meaningful tasks of intermediary complexity, both from low-level and from high-level tasks.

5.6 WG: Novel Visual Metaphors for Multivariate Networks

Oliver Kohlbacher, Jonathan C. Roberts, Matthew O. Ward, Jing Yang, and Michelle X. Zhou

There are opportunities to display multivariate networks in different ways. While node-link representations are very common, there are a lot of other possibilities. We look for novel visual metaphors and visual interaction techniques for multivariate networks and discuss the data visual mapping process involving them. The metaphors are organized into nature inspired, geographical, non-geographical, man-made, non-physical, and traditional-visualization inspired classes. Besides exemplar work from each class, several case studies of prior work on novel metaphors are presented with in-depth discussions on advantages and limitations. Finally, a gallery of new metaphors is presented in the book chapter.

6 Open Problems in Multivariate Network Visualization

Not all the topics identified during the seminar could be addressed in the working groups (and will not be addressed in the planned book) and might be considered for a future Dagstuhl seminar. They include the following:

- Data quality: How can we represent uncertainty, noise, errors, missing data?
- Dimensions: How can we best characterize/reduce the high-dimensional variable space in multivariate networks?
- Evaluation and Experimental Design: What tasks do scientists perform with visualizations? How do we measure understanding of such visualizations? What is the role of aesthetics?
- Existing Practices: How can we merge the best practices of graph drawing and information visualization?
- Scalability: How can we make visualizations of large data sets meaningful?
- Selective Visualization: What should be shown and when?
- Faithfulness in Visualization: How can we ensure the validity of visualizations?
- Toolkits and Standards: What progress has been made to date on achieving toolkits and standard data/information representations?
- Challenges for the Next Ten Years: What challenges can be solved and what challenges do we expect to come up during the next ten years?

7 Acknowledgments

We would like to thank all participants of the seminar for the lively discussions and contributions during the seminar as well as the scientific directorate of Dagstuhl Castle for giving us the possibility of organizing this event. Björn Zimmer gathered the abstracts for the overview of talks in Sect. 4 of this document and the slides of all presenters. These slides can be found on the materials website of the seminar. In addition, many attendees agreed to take notes during the breakout sessions. These notes were the basis for writing this executive summary (incl. Sect. 5) and are also available for download on the Dagstuhl web pages of the seminar. Last but not least, the seminar would not have been possible without the great help of the staff at Dagstuhl Castle. We acknowledge all of them and their assistance.

Participants

- James Abello
Rutgers Univ. – Piscataway, US
- Daniel Archambault
Swansea University, GB
- Katy Börner
Indiana University –
Bloomington, US
- Maura Conway
Dublin City University, IE
- Stephan Diehl
Universität Trier, DE
- Tim Dwyer
Monash Univ. Melbourne, AU
- Peter Eades
The University of Sydney, AU
- Niklas Elmqvist
Purdue University, US
- Jean-Daniel Fekete
INRIA Saclay – Île-de-France –
Orsay, FR
- Helen Gibson
University of Northumbria, GB
- Carsten Görg
University of Colorado, US
- Hans Hagen
TU Kaiserslautern, DE
- Benjamin David Hennig
University of Sheffield, GB
- Danny Holten
SynerScope BV, NL
- Christophe Hurter
ENAC – Toulouse, FR
- T. J. Jankun-Kelly
Mississippi State University, US
- Daniel A. Keim
Universität Konstanz, DE
- Jessie Kennedy
Edinburgh Napier University, GB
- Andreas Kerren
Linnaeus University – Växjö, SE
- Stephen G. Kobourov
Univ. of Arizona – Tucson, US
- Oliver Kohlbacher
Universität Tübingen, DE
- Robert Kosara
Tableau Software – Seattle, US
- Lothar Krempel
MPI für Gesellschaftsforschung –
Köln, DE
- Kwan-Liu Ma
Univ. of California – Davis, US
- Guy Melançon
University of Bordeaux, FR
- Silvia Miksch
TU Wien, AT
- Martin Nöllenburg
KIT – Karlsruhe Institute of
Technology, DE
- A. Johannes Pretorius
University of Leeds, GB
- Helen C. Purchase
University of Glasgow, GB
- Jonathan C. Roberts
Bangor University, GB
- Falk Schreiber
IPK Gatersleben &
MLU Halle, DE
- John T. Stasko
Georgia Inst. of Technology, US
- Alexandru C. Telea
University of Groningen, NL
- Jarke J. van Wijk
TU Eindhoven, NL
- Tatiana von Landesberger
TU Darmstadt, DE
- Matthew O. Ward
Worcester Polytechnic Inst., US
- Michael Wybrow
Monash Univ. Melbourne, AU
- Kai Xu
Middlesex University, GB
- Jing Yang
University of North Carolina at
Charlotte, US
- Michelle X. Zhou
IBM Almaden Center –
San José, US
- Björn Zimmer
Linnaeus University – Växjö, SE

Report from Dagstuhl Seminar 13211

Automated Reasoning on Conceptual Schemas

Edited by

Diego Calvanese[1], Sven Hartmann[2], and Ernest Teniente[3]

1 **Free University Bozen-Bolzano, IT,** calvanese@inf.unibz.it
2 **TU Clausthal, DE,** sven.hartmann@tu-clausthal.de
3 **UPC – Barcelona, ES,** teniente@essi.upc.edu

Abstract

This report documents the outcomes of the Dagstuhl Seminar 13211 "Automated Reasoning on Conceptual Schemas". The quality of an information system is largely determined early in the development cycle, i.e., during requirements specification and conceptual modeling since errors introduced at these stages are usually much more expensive to correct than errors made during design or implementation. Thus, it is desirable to prevent, detect, and correct errors as early as possible in the development process by assessing the correctness of the conceptual schemas built. The high expressivity of conceptual schemas requires to adopt automated reasoning techniques to support the designer in this important task.

Research in this area can be classified according to two different dimensions. On the one hand, according to the language used to specify the conceptual schema. On the other hand, according to whether reasoning is performed on the structural schema alone, or also on its dynamic aspects. We find interesting and promising results from all these communities which have usually worked isolatedly. Therefore, the aim of this seminar was to allow them to communicate with each other to avoid duplicate effort and to exploit synergies. The research questions that were pursued in the seminar included, among others: (i) Does it make sense to renounce to decidability to be able to handle the full expressive power of the language used with and without textual integrity constraints? (ii) Which is the current state of the achievements as far as reasoning on the behavioral part is concerned? (iii) Are the existing techniques and tools ready to be used in an industrial environment? (iv) Which are the new challenges for automated reasoning on conceptual schemas?

Seminar 19–24 May, 2013 – www.dagstuhl.de/13211
1998 ACM Subject Classification D.2.4 [Software Engineering] Software/Program Verification, F.3.1 [Logics and meanings of programs] Specifying, Verifying and Reasoning about Programs
Keywords and phrases Automated Reasoning, Conceptual Schema of an Information System, Validation, Verification
Digital Object Identifier 10.4230/DagRep.3.5.43
Edited in cooperation with Xavier Oriol

1 Executive Summary

Diego Calvanese
Sven Hartmann
Ernest Teniente

This Dagstuhl Seminar brought together 37 researchers from 16 countries across disciplines related to automated reasoning on conceptual schemas. The participants' expertise covered

Automated Reasoning on Conceptual Schemas, *Dagstuhl Reports*, Vol. 3, Issue 5, pp. 43–77
Editors: Diego Calvanese, Sven Hartmann, and Ernest Teniente
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

	Monday	Tuesday	Wednesday	Thursday	Friday
08:30		Reasoning about the Dynamics			
09:00	*Opening* Reasoning on Structural Schemas (I)		Break Out Session	Break Out Session	Group Conclusions
10:00	Coffee Break				
10:30	Reasoning on Structural Schemas (II)	New Challenges	Break Out Session	Break Out Session	Group Conclusions
12:00	Lunch				
14:00	Reasoning on Structural Schemas (III)	Reasoning about Mappings	*Excursion*	Reports of the Break Out Sessions	
15:30	Coffee Break			Coffee Break	
16:00	Extensions	Reasoning about Dependencies		Discussion	
18:00	Dinner				

Figure 1 Timetable of the seminar.

the three most popular languages used to specify the conceptual schema, i.e., Entity-Relationship (ER), Unified Modeling Language (UML) and Object-Role Modeling (ORM); either addressing reasoning only on the static (i.e., structural) schema alone or reasoning also on the elements of a conceptual schema that capture the dynamic (i.e., behavioral) aspects of a system.

Monday and Tuesday were devoted to short presentations from the participants of their most recent achievements in the field.

On Wednesday and Thursday morning the participants were allocated to three different groups, in parallel break out sessions, each one of them addressing a different aspect related to the topic of the workshop:

- On the practical applicability of current techniques for reasoning on the structural schema;
- Reasoning about the conceptual schema components capturing dynamic aspects;
- New challenges for automated reasoning on conceptual schemas.

The organizers asked each group to share the experiences of their participants and to try to identify the most pressing and challenging research issues or open problems for the aspect it addressed. Each group presented a summary of their results on Thursday afternoon. Thursday evening and Friday morning were devoted to a discussion about the outcomes of each group aiming at trying to come up with a roadmap for automated reasoning on conceptual schemas, something which was shown to be harder than expected.

The concrete timetable of the seminar is shown in Figure 1.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

3.1 Reasoning over Conceptual Data Models: The Description Logic Approach

Alessandro Artale (Free University of Bozen-Bolzano, IT)

Joint work of Artale, Alessandro; Calvanese, Diego; Kontchakov, Roman; Ryzhikov, Vladislav; Zakharyaschev, Michael;
Main reference A. Artale, D. Calvanese, R. Kontchakov, V. Ryzhikov, M. Zakharyaschev, "Reasoning over extended ER models," in: Conceptual Modeling – ER 2007, LNCS, Vol. 4801, pp. 277–292, Springer, 2007.
URL http://dx.doi.org/10.1007/978-3-540-75563-0_20

In this talk we show how reasoning techniques developed in the field of knowledge representation (Description Logics, in particular) can be used to check quality properties of Conceptual Models [4], i.e., schema consistency, class consistency, instance checking and model checking [6, 5, 3, 1]. We consider various fragments of the language of Extended Entity-Relationship (EER) diagrams, which includes a number of constructs: ISA between entities and relationships, disjointness and covering of entities and relationships, cardinality constraints for entities in relationships and their refinements as well as multiplicity constraints for attributes.

The main results are obtained by mapping ER constructs to, so called, DL-Lite [7, 2] logics, thus showing the usefulness of such languages for reasoning over conceptual models and ontologies. The talk will also emphasise the difference between the database and the ontology point of view of Conceptual Models.

References

1 A. Artale, D. Calvanese, R. Kontchakov, V. Ryzhikov, and M. Zakharyaschev. Reasoning over extended ER models. In *Proc. of the* 26th *International Conference on Conceptual Modeling (ER'07)*, volume 4801, Auckland, New Zealand, Nov. 2007. Lecture Notes in CS, Springer.

2 A. Artale, D. Calvanese, R. Kontchakov, and M. Zakharyaschev. The DL-Lite family and relations. *Journal of Artificial Intelligence Research (JAIR)*, 36:1–69, 2009.

3 A. Artale, D. Calvanese, and A. Ibáñez-García. Full satisfiability of uml class diagrams. In *Proc. of the* 29th *International Conference on Conceptual Modeling (ER-10)*, 2010.

4 C. Batini, S. Ceri, and S.B. Navathe. *Conceptual Database Design, an Entity-Relationship Approach.* Benjamin and Cummings Publ. Co., 1992.

5 D. Berardi, D. Calvanese, and G. De Giacomo. Reasoning on UML class diagrams. *Artificial Intelligence*, 168(1–2):70–118, 2005.

6 D. Calvanese, M. Lenzerini, and D. Nardi. Unifying class-based representation formalisms. *J. of Artificial Intelligence Research*, 11:199–240, 1999.

7 D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Data complexity of query answering in description logics. In *Proc. of the 10th Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR 2006)*, pages 260–270, 2006.

3.2 UML class diagrams – decision, identification and repair of correctness and quality problems

Mira Balaban (Ben Gurion University – Beer Sheva, IL)

Joint work of Balaban, Mira; Maraee Azzam; Sturm Arnon; Kifer Michael; Khitron Igal
Main reference M. Balaban, A. Maraee, "Finite satisfiability of UML class diagrams with constrained class hierarchy," ACM Trans. on Software Engineering and Methodology (TOSEM), 22(3), 24pp., ACM, 2013.
URL http://dx.doi.org/10.1145/2491509.2491518

We have developed efficient methods for analysis of correctness and quality problems in UML class diagrams. The ultimate goal is to have a rich support for static analysis of models, so to enable the development of advanced model level IDEs. Our methods are implemented in our FiniteSatUSE tool and its associated catalog of correctness anti-patterns. The methods:

1. Detection of Finite Satisfiability problems.
2. Identification of the cause for a Finite Satisfiability or a Consistency problem.
3. Repair – for typical problems. Based on the catalog.
4. Simplification of diagrams – remove redundant constraints.
5. Completion of diagrams – discover hidden constraints.

Most methods apply to subsets of UML class diagrams – depending on included constraints and structure of the diagram. Some methods are complete for certain subsets, but become incomplete when new constraints are added.

Various questions arise from this research:

1. The need for abstraction means in class diagrams – Our Pattern Class Diagram language is a start.
2. Efficient discovery of patterns in a class diagram.
3. Creation of a benchmark of class diagrams, for evaluation of analysis algorithms – Metric directed class diagram constructor; e.g., construct a class diagram that satisfies a structure metric of presence of association- cycles, or a size-metric of proportion of classes and associations.
4. Integration of (1) a UML tool; with (2) a reasoning underlying module (our F-OML project), and with (3) the above methods.

All relevant material is in: http://www.cs.bgu.ac.il/ modeling/

References

1 M. Balaban and A. Maraee *Finite Satisfiability of UML Class Diagrams with Constrained Class Hierarchy*. TOSEM, to appear. ACM (2013).

2 M. Balaban and A. Maraee *Inter-association Constraints in UML2: Comparative Analysis, Usage Recommendations, and Modeling Guidelines*. MoDELS (2012).

3 M. Balaban and M. Kifer *Logic-based Model-Level Software Development with F-OML*. MoDELS (2011).

4 M. Balaban and A. Maraee and A. Sturm *A Pattern-Based Approach for Improving Model Quality*. submitted (2013).

3.3 A Problem Statement: Representation of Instance-Derivations Based on Dependencies

Joachim Biskup (TU Dortmund, DE)

A relational database schema is specified by a set of relation names with attributes and a set of dependencies that constrain the possible instance relations. Regarding the dependencies, mostly two kinds of derivations have been investigated, namely deciding on implications and verifying satisfaction.

In this talk, we will study another kind of derivations, namely: whether and how a user who has an incomplete view on the instance relations can derive further nontrivial details of the instance relations based on the dependencies. More specifically, we would like to obtain a concise representation of all options for such derivations. Our interest is motivated by applications in the field of inference control; it might also be relevant for avoiding redundancy when designing a database.

Intuitively, given the dependencies we would like find a collection of "derivation-skeletons" such that the following properties are achieved: correctness for all instances, completeness for all instances, and neither local nor global redundancy.

Having no substantial results so far, we basically only have the following conjecture: A set of dependencies (having some nice properties like being "full" (not embedded)), admits a finite collection of "derivation-skeletons" with the required properties iff the given set is some sense "acyclic" (still to be made precise).

3.4 Employing Automatic Reasoning on Conceptual Schemas

Joachim Biskup (TU Dortmund, DE)

Joint work of Biskup, Joachim; Hartmann, Sven

We suggest the following classication of how automatic reasoning on conceptual schemas is employed:

1. At design time:
   - prove assurances for given schema:
     * assurances for a single state:

       ...
     * assurances for sequences of states:

       ...
   - transform given schema into "better" one (satisfying more useful assurances)
   - exhibit inherent conflicts/tradeoffs to achieve useful assurances
   - integrate new aspects or independently specified aspects
   - generate "closed" views from specification of informational needs
   - generate "representative" instances for illustrating practical consequences
   - estimate future costs (runtime, space, communication, ...)
2. At run time, continued employment using the assurances proven at design time:
   - improve quality of answers to queries

– supply/suggest missing inputs for updates of data (instances)
– ensure meaningful results of complex data manipulations
– support prediction and optimization of operational requests
– support schema consolidation
– assist in enforcing assurances during evolvement of informational needs
– enable interoperation/communication with other agents, in particular for complex querying

3.5 Incremental inconsistencies detection with low memory overhead

Xavier Blanc (University of Bordeaux, FR)

As ensuring models' consistency is a key concern when using a model-based development approach, model inconsistency detection has received significant attention over the last years. To be useful, inconsistency detection has to be sound, efficient and scalable. Sound means that inconsistencies detected by a tool should be correct, and not mislead the developers in their design tasks. Efficient means that detection has to be executed as fast as possible because developers always want to know the impact of their modification immediately after having performing them. Scalability is however a little more complex as a concern and usually defines specific requirements that are deemed important depending on the context. In this talk, we presented a new incremental inconsistency detection approach that emphasis on scalability as it only consumes a small and model size-independent amount of memory.

3.6 Modeling@SAP: Why Class Models Are Rarely Used

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

In 1999, SAP started to combine the Unified Modeling Language (UML) and the Fundamental Modeling Concepts (FMC) language (for details, see http://www.fmc-modeling.org/fmc-and-tam/). The result is an SAP internal standard for modeling, called Technical Architecture Modeling (TAM). TAM comprises block diagrams, component diagrams, package diagrams, class diagrams, activity diagrams, sequence diagrams, state diagrams, and use case diagrams. TAM is used for both conceptual modeling as well as design modeling.

While many works on reasoning on conceptual schemas focus on class diagrams and state diagrams, the most often used diagram type used for conceptual modeling at SAP is the block diagram. For example, class models are used rarely, as they are "too close to real code." In general, developers and architects prefer high-level structural diagrams (e.g., block diagrams), thus we need to ask ourselves the questions, if we can reason over such models and what kind of properties help to improve the software development.

3.7 Reasoning over Secure Business Processes

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

Enterprise systems are often process-driven. In such systems, high-level process-models play an important role both for communicating business requirements between domain experts and system experts as well as for system implementation. Since several years, enterprise systems need to fulfill an increasing number of the security and compliance requirements. Thus, there is an increasing demand for integrating high-level security and compliance requirements into process models.

In general, we present an approach for reasoning over secure process-models, i.e., process-models containing high-level security and compliance requirements. In particular, this approach helps to detect non-compliant or inconsistent process-models early during both the modeling of business processes as well as during system development.

References

1 A.D. Brucker and I. Hang. Secure and compliant implementation of business process-driven systems. In Marcello La Rosa and Pnina Soffer, editors, *Joint Workshop on Security in Business Processes (SBP)*, volume 132 of *Lecture Notes in Business Information Processing (LNBIP)*, pages 662–674. Springer-Verlag, 2012.

2 A.D. Brucker, I. Hang, G. Lückemeyer, and R. Ruparel. SecureBPMN: Modeling and enforcing access control requirements in business processes. In *ACM symposium on access control models and technologies (SACMAT)*, pages 123–126. ACM Press, 2012.

3 L. Compagna, P. Guilleminot, and A.D. Brucker. Business process compliance via security validation as a service. In Manuel Oriol and John Penix, editors, *Testing Tools Track of International Conference on Software Testing, Verification, and Validation (Tools@ICST)*. IEEE Computer Society, 2013.

3.8 Modeling and Reasoning over Processes and Data

Diego Calvanese (Free University of Bozen-Bolzano, IT)

Joint work of Calvanese, Diego; Bagheri Hariri, Babak; Montali, Marco; Santoso, Ario; De Giacomo, Giuseppe; Deutsch, Alin
Main reference B. Bagheri Hariri, D. Calvanese, G. De Giacomo, A. Deutsch, M. Montali, "Verification of relational data-centric dynamic systems with external services," arXiv:1203.0024v1 [cs.DB], 2012.
URL http://arxiv.org/abs/1203.0024v1

Data and processes are just two sides of the same coin, and for several activities related to the analysis and design of systems it is therefore important to capture both static and dynamic aspects in a uniform way. Data-centric dynamic systems (DCDSs) are systems where both the process controlling the dynamics and the manipulation of data are equally central, and are captured in a pristine way, abstracting from specific features of concrete models. DCDSs allow one to capture commonly adopted models for data and processes, such as artifact systems modeled as finite state machines or through the Guard-Stage- Milestone (GSM) model. Hence, they are well suited for the formal analysis and verification of temporal properties over the evolution of such systems. In the talk we present DCDSs and discuss the results on decidadibility of verification of first-order variants of μ-calculus over such systems. We also point to extensions of DCDSs with a semantic level, allowing one to capture properties at a higher level of abstraction, and taking into account semantic constraints during verification.

3.9 Preliminary Report on an Algebra of Lightweight Ontologies

Marco A. Casanova (PUC – Rio de Janeiro, BR)

Joint work of Casanova, Marco A.; Sacramento, Eveline R.; Macêdo, José A. F.: Pinheiro, Ângela M. A.; Vidal, Vânia M. P.; Breitman, Karin K.; Furtado, Antonio L.
Main reference M.A. Casanova, J.A.F. de Macêdo, E.R. Sacramento, Â.M.A. Pinheiro, V.M.P. Vidal, K.K. Breitman, A.L. Furtado, "Operations over lightweight ontologies," in: On the Move to Meaningful Internet Systems (OTM'12), LNCS, Vol. 7566, pp. 646–663, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-33615-7_14

We argue that certain familiar ontology design problems are profitably addressed by treating ontologies as theories and by defining a set of operations that create new ontologies, including their constraints, out of other ontologies. Such operations extend the idea of namespaces to take into account constraints.

Consider first the problem of designing an ontology to publish data on the Web. If the designer follows the Linked Data principles, he must select known ontologies, as much as possible, to organize the data so that applications can dereference the URIs that identify vocabulary terms in order to find their definition. We argue that the designer should go further and analyze the constraints of the ontologies from which he is drawing the terms to construct his vocabulary. Furthermore, he should publish the data in such a way that the original semantics of the terms is preserved. To facilitate ontology design from this perspective, we introduce three operations on ontologies, called projection, union and deprecation.

Consider now the problem of comparing the expressive power of two ontologies, O = (V, Σ) and O' = (V', Σ'). If the designer wants to know what they have in common, he should create a mapping between their vocabularies and detect which constraints hold in both ontologies, after the terms are appropriately mapped. The intersection operation answers this question. We argued elsewhere (Casanova et al., 2010) that intersection is also useful to address the design of mediated schemas that combine several export schemas in a way that the data exposed by the mediator is always consistent.

On the other hand, if the designer wants to know what holds in O = (V, Σ), but not in O' = (V', Σ'), he should again create a mapping between their vocabularies and detect which constraints hold in the theory of Σ, but not in the theory of Σ', after the terms are appropriately mapped. The difference operation answers this question.

3.10 OCL2FOL: Using SMT solvers to automatically reason on conceptual schemata with OCL constraints

Carolina Dania (IMDEA Software Institute, ES)

In this talk we present a mapping from a rich subset of OCL expressions into first order logic, and discuss how this mapping can be used to automatically check, using SMT solvers, the satisfiability of conceptual schemata with OCL constraints. We also present our OCL2FOL tool, which implements our mapping and generates, for each conceptual schema and OCL constraints, the corresponding satisfiability problem for Z3 or Yices.

3.11 Metrics for visual notations

Sophie Dupuy-Chessa (LIG – Grenoble, FR)

Joint work of Le Pallec, Xavier; Mandran, Nadine; Genon Nicolas

The maturity of Model Driven Engineering facilitates the development of domain specific languages. The creation of the languages relies on the definition of metamodels, but also on their corresponding visual notations. But one can wonder what is the quality of a new language. It can result in inunderstandable diagrams with inappropriate notations. Then our goal is to evaluate the quality of metamodels and notations by metrics, which are integrated in a (meta)modeling environment. In particular, we are studying metrics for visual notations in order to automate the measure of quality for a notation and its derived diagrams.

3.12 "Automating Reasoning on Conceptual Schemas" in FamilySearch^{TM} – a Large-Scale Reasoning Application

David W. Embley (Brigham Young University, US)

Among its many other projects, FamilySearch.org has scanned and OCRed 85,000 books filled with family-history information—an estimated 4.25×10^{12} "facts" of interest. We wish to extract and organize these facts for semantic search with respect to a conceptual model. Reasoning applies in several ways: (1) traditional satisfiability checks over potentially more expressive constraints, (2) inference with large sets of assertions—about half of the facts are inferred; (3) constraint violation search —unlike many database applications, it's almost certain that the facts will not satisfy the constraints of the conceptual model, so finding the myriad of discrepancies is of interest; and (4) uncertainty —both facts and constraints are uncertain, so probabilistic description logics are of interest. The presentation will likely have more questions than answers, and in that sense, could lead to interesting discussion and possible directions for future work.

3.13 Constraints on Class Diagrams

Ingo Feinerer (TU Wien, AT)

We give a short overview of our approach for encoding UML class diagrams for checking satisfiability and for computing minimal instances. We aim for efficient and lightweight techniques, like integer linear programming and flow networks, that scale well also in industrial-scale application domains. As open problems for discussion we highlight a few constraints, typically expressed via OCL, that we found relevant in industrial settings but which are still challenging from a research perspective, like association chains or equations over them.

3.14 ORM2: formalisation and encoding in OWL2

Enrico Franconi (Free University of Bozen-Bolzano, IT)

Joint work of Franconi, Enrico; Mosca, Alessandro
Main reference E. Franconi, A. Mosca, D. Solomakhin, "ORM2: Formalisation and encoding in OWL2," in: On the Move to Meaningful Internet Systems (OTM'12), LNCS, Vol. 7567, pp. 368–378, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-33618-8_51

The introduction of a provably correct encoding of a fragment of ORM2 (called ORM2zero) into a decidable fragment of OWL2, opened the doors for the definition of dedicated reasoning technologies supporting the quality of the schemas design.

3.15 Reasoning over Order Dependencies for Relational Schema

Parke Godfrey (York University – Toronto, CA)

Dependency theory has played important roles for relational databases, with functional dependencies (FDs) as a hallmark. Structural constraints provide a framework for formalizing good design, as by normalization and decomposition, and for reasoning over schemas. Much has been studied on how to reason effectively over dependencies. Armstrong's Axioms provide a sound and complete axiomatization for FDs, which led to insights for efficient inference procedures. Such "reasoners" are easier to understand and prove much more efficient that general reasoning from first principles.

Dependencies have also played a critical role on the physical side of database systems. They are used within query optimization extensively. For instance, FDs can be used to convert one group-by specification as stated within an SQL query into a logically equivalent one, but one more amenable to an efficient query plan. Thus, query optimizers incorporate limited reasoners for dependencies.

Order dependencies (ODs) generalize functional dependencies. ODs are to FDs as order by is to group by. An OD tells that when data is sorted with respect to one specification, it then must also be sorted with respect to another. We have been studying ODs in depth recently, and we know much more about them now. We have devised a sound and complete axiomatization for them as Armstrong's Axioms are for FDs. From this, we are developing efficient inference procedures for them. We know much about the complexities of inference tasks for ODs. We have shown a proper hierarchy for OD classes. Our motivation for this endeavor has been, in large part, to use ODs effectively within query optimization. ODs offer a powerful means to exchange one interesting order on a data stream (within the runtime of the query plan) for another (when it is known the second implies the first), which broadens the space of possible plans.

However, ODs could also play a vital role in design, as do FDs. We do not understand yet how ODs can be viewed as structural constraints, and what roles they might play in schemas and reasoning over schemas. This is a worthwhile topic to explore.

3.16 Exploring UML and OCL Model Properties with Relational Logic

Martin Gogolla (Universität Bremen, DE)

Joint work of USE Development Team
Main reference M. Kuhlmann, M. Gogolla, "From UML and OCL to relational logic and back," in: Model Driven Engineering Languages and Systems, LNCS, Vol. 7590, pp. 415–431, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-33666-9_27
URL http://sourceforge.net/apps/mediawiki/useocl/index.php

Modeling Languages like UML or EMF employ OCL in order to precisely model systems in terms of a conceptual schema for applications. Complex UML and OCL models therefore represent a crucial part of model-based engineering, as they allow to formally describe central system properties like model consistency or constraint independence, among other important properties. We discuss a lightweight model validation and verification method based on efficient SAT solving techniques. Our work relies on a transformation from UML class diagrams and OCL concepts into relational logic. Relational logic in turn represents the source for advanced SAT-based model finders. The approach allows us to explicitly benefit from the efficient handling of relational logic and to interpret found results backwards in terms of UML and OCL. We explain our ideas with an example from the original work by Peter Chen on conceptual modeling with the entity-relationship model.

References

1 M. Kuhlmann and M. Gogolla. From UML and OCL to Relational Logic and Back. In Robert France, Juergen Kazmeier, Ruth Breu, and Colin Atkinson, editors, *Proc. 15th Int. Conf. Model Driven Engineering Languages and Systems (MoDELS'2012)*, pages 415-431. Springer, Berlin, LNCS 7590, 2012.

3.17 Armstrong Instances as a Reasoning Aid

Sven Hartmann (TU Clausthal, DE)

Joint work of Hartmann, Sven; Link, Sebastian; Ferrarotti, Flavio; Köhler, Henning; Le, Van; Leck, Uwe; Thalheim, Bernhard; Trinh, Thu

In our talk we survey recent results on the structure, existence and computation of Armstrong instances for general cardinality constraints and functional dependencies in partial databases. Leading database design tools recommend the creation of data samples to validate, communicate, and evolve the database models they produce. Armstrong instances are perfect data samples in the sense that they satisfy exactly those integrity constraints within a particular constraint class that are logical consequences of the specified constraint set. They can be queried by the data engineer to test the implication of constraints within the constraint class of interest. We discuss characterizations and constructions of Armstrong instances, and list open problems that we suggest for future research.

3.18 Automated Design of Updateable Database Views: a Framework for Possible Strategies

Stephen J. Hegner (University of Umeå, SE)

There has been substantial amount of work on the subject of how updates to database views should be supported. However, there has been little reported on how to design views which meet certain requirements for information content and furthermore support a certain set U of updates. In this work, some ideas on how to automate the design of views which are updateable via the constant-meet-complement strategy are presented. The design process itself may require some flexibility in the choice of the view to be updated, since a larger view will generally admit a larger set of updates. Thus, the process consists of the identification of a pair of views (Γ,Γ'), in which Γ is the view to be updated and Γ' is a meet complement which supports all updates in U. The constraint on Γ is that it recapture a certain prespecified minimal amount of information $I_{\min}$ from the main schema, but not more than a prespecified upper bound $I_{\max}$. Since meet complements are precisely those which admit an embedded cover of the dependencies of the main schema, the key to an effective realization of this approach is to find methods for the efficient identification of embedded covers.

3.19 An ontology-driven unifying metamodel for UML Class Diagrams, EER, and ORM2

C. Maria Keet (University of KwaZulu-Natal – Durban, ZA)

Joint work of Keet, C. Maria; Fillottrani, Pablo Rubén

Software interoperability may be achieved by using their respective conceptual data models. However, each model may be represented in a different conceptual data modelling language for the tool's purpose or due to legacy issues. Several translations between small subsets of language features are known, but no unified model exists that includes all their language features. Aiming toward filling this gap, we designed a common and unified, ontology-driven, metamodel covering and unifying EER, UML Class Diagrams v2.4.1, and ORM2. We present the static, structural, components of the metamodel, highlighting the common entities and summarizing some modelling motivations. This metamodel will be taken as the basis for a comprehensive formalization, which afterward may be used for reasoning over the structural components of individual and linked conceptual models represented in any of UML, EER, or ORM2. More details of the project and the related papers are available online at http://www.meteck.org/SAAR.html

References

1 C.M. Keet, P.R. Fillottrani. Toward an ontology-driven unifying metamodel for UML Class Diagrams, EER, and ORM2. *32nd International Conference on Conceptual Modeling (ER'13)*. 11-13 November, 2013, Hong Kong. Springer LNCS (accepted).

2 C.M. Keet, P.R. Fillottrani. Structural entities of an ontology-driven unifying metamodel for UML, EER, and ORM2. *3rd International Conference on Model and Data Engineering (MEDI'13)*. September 25-27, 2013, Amantea, Calabria, Italy. Springer LNCS (accepted).

3.20 Temporal Extended Conceptual Models

Roman Kontchakov (Birkbeck College London, GB)

In this talk we discuss the temporal description logics [3], designed for reasoning about temporal conceptual data models, and investigate their computational complexity. Our formalisms are based on the DL-Lite family of description logics with three types of concept inclusions (ranging from atomic concept inclusions and disjointness to the full Booleans), as well as cardinality constraints and role inclusions [4]. In the temporal dimension, they capture future and past temporal operators on concepts, flexible and rigid roles, the operators 'always' and 'some time' on roles, data assertions for particular moments of time and global concept inclusions. The logics are interpreted over the Cartesian products of object domains and the flow of time (Z,<), satisfying the constant domain assumption. We prove [2] that the most expressive of our temporal description logics (which can capture lifespan cardinalities and either qualitative or quantitative evolution constraints) turn out to be undecidable. However, by omitting some of the temporal operators on concepts/roles or by restricting the form of concept inclusions we obtain logics whose complexity ranges between PSpace and NLogSpace. These positive and encouraging results were obtained by reduction to various clausal fragments of propositional temporal logic, which opens a way to employ propositional or first-order temporal provers for reasoning about temporal data models.

In the second part of the talk we focus on ontology-based data access (OBDA), where the most important technique for answering queries is to rewrite the given ontology and the query into a single first-order query that can be evaluated directly over the data. The current W3C standard for OBDA is OWL 2 QL which is based on the DL-Lite family of description logics. Our aim is to extend standard atemporal OBDA to data with validity time and ontologies that are suitable for temporal conceptual modelling. To this end, we design a temporal description logic, TQL, that extends the standard ontology language OWL 2 QL, provides basic means for temporal conceptual modelling and ensures first-order rewritability of conjunctive queries for suitably defined data instances with validity time [1].

References

1 A. Artale, R. Kontchakov, F. Wolter and M. Zakharyaschev. Temporal Description Logic for Ontology-Based Data Access. In *Proceedings of IJCAI* (Beijing, 3-9 August), 2013. full version is available at arXiv:1304.5185

2 A. Artale, R. Kontchakov, V. Ryzhikov and M. Zakharyaschev. A Cookbook for Temporal Conceptual Data Modelling with Description Logics, 2012. arXiv:1209.5571.

3 A. Artale, R. Kontchakov, V. Ryzhikov and M. Zakharyaschev. Tailoring Temporal Description Logics for Reasoning over Temporal Conceptual Models. In *Proc. of the 8th Inter-*

national Symposium on Frontiers of Combining Systems (FroCoS 2011). LNCS, vol. 6989, pages 1-11. Springer, 2011.

4 A. Artale, D. Calvanese, R. Kontchakov and M. Zakharyaschev. The DL-Lite Family and Relations. J. Artif. Intell. Res. (JAIR), 36:1-69, 2009.

3.21 ProB: Solving Constraints on Large Data and Higher-Order Formal Models

Michael Leuschel (Heinrich-Heine-Universität Düsseldorf, DE))

The B-method is a formal method for specifying safety critical systems, reasoning about those systems and generate code that is correct by construction. B is based on predicate logic, augmented with arithmetic (over integers), (typed) set theory, as well as operators for relations, functions and sequences. As such, B provides a very expressive foundation which is familiar to many mathematicians and computer scientists. One of our goals is to be able to use B as a high-level and easy to use language to express constraints in wide variety of application areas. In this talk we have presented the ProB [1] constraint solver and model checker for the B method. In particular, we describe how a variety of other formalisms and industrial problems can be mapped to B and how ProB can be used for automated reasoning. For example, we have discussed the iUML tool and its integration into the Rodin toolset and link with ProB for animation and validation. Finally, we also touch upon the various backends [2] of ProB that can be used in practice.

References

1 D. Plagge and M. Leuschel. Seven at a stroke: LTL model checking for high-level specifications in B, Z, CSP, and more. In *International journal on software tools for technology transfer*. vol. 12, pages 9-21. Springer, 2010.

2 D. Plagge and M. Leuschel. Validating B, Z and TLA+ using ProB and Kodkod. In D. Giannakopoulou and D. Méry, editors, *Proceedings FM'2012*, LNCS 7436, pages 372-386. Springer, 2012.

3.22 A Declarative Approach to Distributed Computing

Jorge Lobo (Universitat Pompeu Fabra – Barcelona, ES)

Joint work of Ma, Jiefei; Le, Franck; Wood, David; Russo, Alessandra; Lobo, Jorge
Main reference A Declarative Approach to Distributed Computing: Specification, Execution and Analysis. To appear in Theory and Practice of Logic Porgramming.

There is an increasing interest in using logic programming to specify and implement distributed algorithms, including a variety of network applications. These are applications where data and computation are distributed among several devices and where, in principle, all the devices can exchange data and share the computational results of the group. In this paper we propose a declarative approach to distributed computing whereby distributed algorithms and communication models can be (i) specified as action theories of uents and actions;

(ii) executed as collections of distributed state machines, where devices are abstracted as (input/output) automata that can exchange messages; and (iii) analysed using existing results on connecting causal theories and Answer Set Programming. Results on the application of our approach to different classes of network protocols are also presented.

3.23 Reasoning on conceptual schemas of spatial data

Stephan Maes (TU Dresden, DE)

Main reference S. Maes, "Reasoning on class relations: An overview," in: Cognitive and Linguistic Aspects of Geographic Space, Lecture Notes in Geoinformation and Cartography, pp. 237–257, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-34359-9_13

Spatial databases require specific integrity constraints to specify restrictions on geometric and topological relations and properties. This work focuses on integrity constraints that define cardinality restrictions for a certain instance relation (for example a topological relation) between all entities of the involved classes. These constraints also allow for automated reasoning to find contradictions and redundancies and to evaluate the compliance with application requirements. Therefore the logical properties of the applied instance relations and those of the cardinality restrictions have to be considered, in particular symmetry and compositions, but also other inferences. The presentation provides an overview of the different types of integrity constraints of spatial data, summarizes research on corresponding reasoning approaches and outlines possible future work. Most of the discussed reasoning algorithms have been implemented in a research prototype that is available at http://www.stephanmaes.de/classrelations.html .

References

1 S. Mäs, . Reasoning on class relations: an overview. In *Raubal, M., Mark, D. M.,and Frank, A. U., editors, Cognitive and Linguistic Aspects of Geographic Space – New Perspectives on Geographic Information Research, Lecture Notes in Geoinformation and Cartography*, pages 237-257. Springer, 2013.

2 S. Mäs, and W. Reinhardt. Categories of geospatial and temporal integrity constraints. In *International Conference on Advanced Geographic Information Systems and Web Services, GEOWS 2009*, pages 146-151. IEEE Computer Society, 2009.

3 S. Mäs. Reasoning on spatial relations between entity classes. In *Cova, T. J., Miller, H. J., Beard, K., Frank, A. U., and Goodchild, M. F., editors, Geographic Information Science, 5th International Conference, GIScience 2008, Proceedings*, volume 5266 of Lecture Notes in Computer Science, pages 234-248. Springer, 2008.

3.24 On BDD, Finite controllability and the BDD/FC conjecture

Jerzy Marcinkowski (University of Wroclaw, PL)

Main reference T. Gogacz, J. Marcinkowski, J., "On the BDD/FC conjecture," in Proc. of the 32nd Symposium on Principles of Database Systems (PODS '13), pp. 127–138, ACM, 2013.
URL http://dx.doi.org/10.1145/2463664.2463668

FC (Finite controllability) and BDD (the Bounded Derivation Depth property) are two properties of Datalog/TGD programs (or of TBoxes, if this is how you wish to call them).

BDD is equivalent to Positive First Order rewritability – the very useful property that allows us to use (all the optimizations of) DBMS in order to compute the certain answers to queries in the presence of a theory.

Finite Controllability of a theory/TBox T means that if the certain answer to a query Q, for a database instance/ ABox D , in the presence of T is 'no' then this 'no' is never a result of an unnatural assumption that the counterexample can be infinite.

We conjecture that for any theory T the property BDD implies FC. We prove this conjecture for the case of binary signatures (which in particular means that it holds true for the Description Logics scenario).

3.25 On the Relationship Between OBDA and Relational Mapping

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

A position talk highlighting connections, similarities and differences between the framework of Ontology-Based Data Access, and the one of Object-Relational Mapping. Despite some key differences, such as the fact that OBDA works with incomplete information and under the assumption that there is an impedance mismatch between the conceptual model and the underlying database, while Object-Relational Mapping works with complete information and assuming a lossless connection between these two layers, we advocate the need for cross-fertilization between the two settings.

3.26 Reasoning about the Effect of Structural Events in UML Conceptual Schemas

Xavier Oriol (UPC – Barcelona, ES)

Joint work of Oriol, Xavier; Teniente, Ernest; Tort, Albert

In this talk, we propose an approach which is aimed at providing feedback regarding the dynamic behaviour of the schema when some set of structural events (i.e. insertion or deletion of instances) happens simultaneously in a consistent information base state.

In this way, given a conceptual schema and an IB state for that schema, we answer how can we insert or delete concrete instances of classes/associations without violating any integrity constraint. Concretely, we answer the minimal sets of structural events that, combined with the desired insertions/deletions, lead the IB state to a new consistent one.

3.27 Information and dependency preserving BCNF decomposition algorithm via attribute splitting

Elena V. Ravve (ORT Braude College – Karmiel, IL)

Joint work of Ravve, Elena V.; Makowsky Johann A.
Main reference J. Makowsky, E. Ravve, "BCNF via attribute splitting," in: Conceptual Modelling and Its Theoretical Foundations, LNCS, Vol. 7260, pp. 73–84, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-28279-9_7

Databases are designed in an interactive way between the database designer and his client. Application driven design uses the language of Entity-Relationship modeling. Another approach consists in collecting the attributes U of all the application, and requires from the client that he specifies the functional dependencies F holding between those attributes. After several iterations this results in a big relation R[U] and a set F of functional dependencies. In the remaining design process, R[U] is decomposed using criteria such as avoiding null values, insertion, deletion and modification anomalies, redundancies, and achieving optimal storage, while keeping U fixed. Boyce-Codd-Heath Normal Form for relational databases was introduced to formulate all these properties in terms of F and the key dependencies derivable from F. We add one more characterization in terms of hidden bijections. We also note that for certain standardized translations of the Entity-Relationship Model into the relational model, the resulting relation schemes are always in BCNF. The classical approach to design is based on iteratively decomposing R[U] using projection, while keeping U fixed and preserving information and the functional dependencies. Unfortunately, BCNF cannot always be achieved in this way. As a compromise 3NF was formulated, which can always be achieved, while preserving information and the functional dependencies. However, not all the FD's follow from the key dependencies. We introduce an additional way of restructuring databases by splitting attributes: The relation scheme is expanded by new attributes, whose interpretation is in bijection with previous attributes. The latter can be expressed using functional dependencies between new and old attributes. The expanded relation scheme then is decomposed into BCNF, preserving information and dependencies up to renaming. Design theory for relational databases fell out of fashion twenty years ago and few papers were published on the topic. However, with the fashionable trend of XML-driven databases, renewed interest in normal forms emerged. Hopefully, our approach can be used to formulate design criteria for XML based databases.

3.28 Semantic-based Mappings

Guillem Rull (UPC – Barcelona, ES)

Joint work of Mecca, Giansalvatore; Rull, Guillem; Santoro, Donatello; Teniente, Ernest
Main reference G. Mecca, G. Rull, D. Santoro, E. Teniente, "Semantic-Based Mappings," in Proc. of the 32nd Int'l Conf. on Conceptual Modeling (ER'13), pp. 11–13, Hong Kong, November, 2013.
URL http://www.essi.upc.edu/~grull/papers/MRST13.pdf

In classical data exchange, where data is to be transferred from a source into a target, schema mappings are specified at the database level. However, in many cases, there is a conceptual schema available for the target database, so we propose to map the source into the target conceptual schema instead of the target database. In this way, the abstraction on the details of how the data is stored in the database provided by the conceptual schema allows for a simpler mapping, which, in turn, makes easier the mapping design task. The challenge is that mappings that target a conceptual schema are not directly executable under the current techniques, so we propose an automatic rewriting algorithm that transforms the given database-to-conceptual schema mapping into a classical database-to-database one.

3.29 The Curse of Restructuring in Dependency Theory

Klaus-Dieter Schewe (Software Competence Center – Hagenberg, AT)

Joint work of Schewe, Klaus-Dieter; Attila, Sali
Main reference A. Sali, K.-D. Schewe, "Weak Functional Dependencies on Trees with Restructuring," Acta Cybern. 20(2), pp. 285–329, 2011.
URL http://www.inf.u-szeged.hu/actacybernetica/edb/vol20n2/Sali_2011_ActaCybernetica.xml

The talk addresses the problem that restructuring in complex value databases (used as a broad term to capture any post-relational database structures) is unavoidable, in particular, if disjoint union is permitted as a constructor. However, the consequences for dependency theory are dramatic. For instance, for (weak) functional dependencies on database structures built from record, lists, sets and multisets a nice axiomatisation can be found, whereas the addition of the union constructor leads to a vast amount of additional structural axioms for wFDs and even non-axiomatisability for FDs. However, the core of the axiomatisation remains the same in the sense that the additional axioms merely cover the properties of coincidence ideals, i.e. the set of subattributes, on which two complex values coincide. This raises the open question, whether it is possible to parameterise dependency theory by isolating such structural properties.

3.30 AuRUS: Automated Reasoning on UML Schemas

Ernest Teniente (UPC – Barcelona, ES)

Joint work of Rull, Guillem; Farré, Carles; Queralt, Anna; Urpí, Toni
Main reference G. Rull, C. Farré, A. Queralt, E. Teniente, T. Urpí, "Aurus: explaining the validation of UML/OCL conceptual schemas," Software and Systems Modeling, 28pp., 2013.
URL http://dx.doi.org/10.1007/s10270-013-0350-8

The validation and the verification of conceptual schemas have attracted a lot of interest during the last years, and several tools have been developed to automate this process as much as possible. This is achieved, in general, by assessing whether the schema satisfies different kinds of desirable properties which ensure that the schema is correct. In this talk we describe AuRUS, a tool we have developed to analyze UML/OCL conceptual schemas and to explain their (in)correctness. When a property is satisfied, AuRUS provides a sample instantiation of the schema showing a particular situation where the property holds. When it is not, AuRUS provides an explanation for such unsatisfiability, i.e., a set of integrity constraints which is in contradiction with the property.

3.31 Visual reasoning with (functional) dependencies

Bernhard Thalheim (Universität Kiel, DE)

Joint work of Bernhard Thalheim; Ove Sörensen
Main reference O. Sörensen, B. Thalheim, "Semantics and pragmatics of integrity constraints," in: Semantics in Data and Knowledge Bases, LNCS, Vol. 7693, pp. 1–17, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-36008-4_1

Logical reasoning is main basis for axiomatisation of constraints. Often first-order predicate calculus is used. Human reasoning is however often spatial. Any child knows the meaning of notions such as "behind", "far". Graphical presentation is used for reasoning and explanation before we learn to write. Most classes of database constraints are however expressed by logical formulas. Typical simple calculi have been developed for functional dependencies or inclusion constraints. The axiomatisation for multivalued dependencies is more difficult to understand and to use. Already the class of cardinality constraints demonstrates that numerical calculi support reasoning far simpler.

We show however that graphical reasoning is far simpler for most classes of database constraints. For instance, functional dependencies can be represented by directed graphs. These graphs allow to reason on sets of functional dependencies. The classical Armstrong axiomatisation can be extended to such graphs. These graphs allow to reason on functional and on negated functional dependencies.

These graphs can also be used for constraint acquisition. It is possible to reason on functional, negated functional and underivable functional dependencies in a simple and powerful fashion. Since normalisation assumes that all valid functional dependencies are known and this problem is exponential of the set of attributes we need more sophisticated approaches.

3.32 Validation of Complex Domain-Specific Modeling Languages

Dániel Varró (Budapest Univ. of Technology & Economics, HU)

Despite the wide range of existing generative tool support, constructing a design environment for a complex domain-specific language (DSL) is still a tedious task as the large number of derived features and well-formedness constraints complementing the domain metamodel necessitate special handling. Incremental model queries as provided by the EMF-IncQuery framework can (i) uniformly specify derived features and well-formedness constraints and (ii) automatically refresh their result set upon model changes. However, for complex domains, such as avionics or automotive, derived features and constraints can be formalized incorrectly resulting in incomplete, ambiguous or inconsistent DSL specifications.

To detect such issues, we propose an automated mapping of EMF metamodels enriched with derived features and well-formedness constraints captured as graph queries in EMF-IncQuery into an effectively propositional fragment of first-order logic which can be efficiently analyzed by the Z3 SMT-solver. Moreover, overapproximations are proposed for complex query features (like transitive closure and recursive calls). Our approach will be illustrated on analyzing DSL being developed for the avionics domain.

We further aim to address to reason about the evolution of models (along potentially infinite state spaces by a combined validation technique based upon shape analysis). Concrete graph based models are abstracted into different kind of shapes to represent the context of model elements in an abstract way. Formal validation is supported on the level of shapes is by a combined use of (a) SMT-solvers to derive potentially relevant context of shape elements, (b) incremental instance-level model queries to filter irrelevant context and (c) some dedicated abstract libraries.

3.33 Reasoning about Dependencies in Schema Mappings

Qing Wang (Australian National University – Canberra, AU)

Schema mappings have been extensively studied in the past decades from a variety of aspects, including high-level specification languages, computational properties, optimization, etc. To discover logical consequences among source constraints, target constraints and mapping constraints of a schema mapping, we develop a graph-based framework that captures the inter-relationship between attributes of different relations. User feedback can be incorporated into the framework to minimize ambiguity in designing schema mappings. In doing so, we can characterize the class of sources instances that have a corresponding target instance under a given schema mapping, and conversely can also capture properties that all target instances for a given source instance must have.

4 Working Groups

4.1 On the Practical Applicability of Current Techniques for Reasoning on the Structural Schema

Ernest Teniente

There has been plenty of promising results for providing automated reasoning on the structural part of the conceptual schema and several prototype tools have been developed with this purpose. However, most of these results have remained at the academical level and the industry is not aware of them or it does not consider them relevant enough since it is not using them in software development. With the aim of reducing the gap between academy and industry, the discussion of the participants in this group was aimed at providing an answer to the following questions:

1. What do we need to convince the industry that this technology is useful?
2. Can we come up with a common vocabulary for the various research disciplines that work on this topic?
3. Can we come up with a research agenda of the problems we have to solve?

The first part of the discussion was devoted to identify the most relevant topics that should be addressed to be able to provide an answer to those questions. In particular, there was an agreement on five different topics: *identifying the relevant properties*, *coming up with a common agreement on the formalization (i.e. definition) of the properties*, *the need for explanations*, *the need for benchmarks*, and *showing the scalability of the tools developed so far*. The second part of the discussion went into digging down for each topic and trying to identify the most relevant issues that should require a proper answer to make the results in this area applicable in practice.

The outcome of the discussions for each topic is summarized in the following:

Properties

Two different kinds of properties were identified: those related to reasoning only at the schema level and properties involving both the schema and the data. The first kind of properties are aimed at detecting whether the schema being defined is correct. So, whenever one such property is not satisfied by the schema, or its results do not correspond to the ones expected by the designer, it means that the schema is not properly defined and it must be necessarily changed. The second kind of properties, i.e. those involving data, should be understood more as *services* provided by the system at run-time rather than properties denoting that the schema is ill-specified at design-time. In general, all the properties arising from the discussions are well-known problems in the area. The most relevant properties for each group are the following:

1. Properties related only to the schema
   - Schema satisfiability. There was an agreement that the empty state should not be accepted as a solution. It was also clear the need to distinguish between finite and infinite satisfiability.
   - Class and association satisfiability

    - Constraints and class redundancy, in the sense that they are entailed by the rest of the schema.
    - User-defined property verification, aimed at allowing the designer to determine whether the state satisfies the requirements of the domain. One possible way to achieve it is by showing the satisfiability of a partially specified state envisaged by the designer.

2. Services involving data
    - *Model checking*, i.e. whether a set of instances satisfies a set of constraints (aka integrity checking). This should be combined with techniques to "restore" consistency when the constraints are violated (or to handle with the "inconsistent" data). It is also worth noting that the database view of data should be taken for this purpose, i.e. closed world assumption: the data is a model of the constraints
    - Satisfiability checking over the data. A special effort should be devoted to deal with incomplete databases (e.g. nulls or disjunctive values). Model generation (aka integrity maintenance) is a must. In this case, the open world assumption is needed in the sense that you can invent new things.
    - Query processing
    - View updating
    - Materialized view maintenance
    - Impact analysis of an update
    - Test-data generation

Definition of the properties

There was an agreement that one of the difficulties with convincing the industry about the usefulness of these properties relies on the lack of agreement in the literature about the precise definition of most of such properties. Therefore, one of the first things the community should do is to agree with their formal definition. There was not enough time for doing this during the break out sessions but the working group identified some issues to be taken into account:

- There is a need to clarify whether finite or infinite interpretations are considered
- There is a need to clarify whether the definition uses the database view or the "open-world" view
- There is also a need to clarify which is the semantics used

Explanations

Having tools to show that all of this works was considered to be the most important general concern for showing that the previous properties can be useful in practice. In addition to being able to check these properties, these tools should also explain the results of performing automated reasoning on the conceptual schema. Specifically, it would be interesting to know what kind of explanations would the industry like to have, what is an explanation and in which language should they be shown to the designer. Moreover, explanations should abstract away from whatever logic is used underneath and they should be given regarding to the model the user is referred to. Facilities for what-if scenarios could also be interesting for the industry. The working group identified also some possible kinds of explanations, making a distinction when the property under validation is satisfied or it is not.

- The property is satisfied
    - Instance or snapshot or witness (generated example)

  - An abstraction of the proof (wrt the terminology of the model)
- The property is not satisfied
  - Providing the (minimal) set(s) of constraints that give raise to the violation
  - An abstraction of the proof (wrt the terminology of the model)
  - Causal reasoning, i.e. suggestions about how to repair the violation

Benchmarks

Benchmarks are very important for industry. However, little attention has been paid to them in the area. In fact, there is not yet an agreement on what a benchmark for automated reasoning on conceptual schemas should be. The working group considered that such benchmarks should at least include a motivation underlying the benchmark (i.e., why is this benchmark for); the schema (and the data, if necessary) under consideration; the properties to be checked by the benchmark and their expected output. The definition language of the benchmark (EER, UML/OCL, ORM, etc.) and the semantics used in the benchmark should also be clearly stated. Additionaly, some questions and ideas arised as a result of the discussions:

- How many benchmarks do we need?
  - It depends on the purpose of the benchmark. Scalability vs language expressivity, for instance.
  - Could we come up with a repository of benchmarks?
- Benchmarks for education seem interesting
- Establishing fair benchmarks
  - Separation of concerns and avoiding conflict of interests are important issues
  - Having a contest for this community could be interesting
- Evaluation criteria are also needed

Scalability

There was a clear agreement that scalability has to be necessarily addressed to convince the industry. Moreover, there seemed to be a common understanding that it would probably be already covered if properties and benchmarks were correctly defined. Other aspects arising from the short discussion we had on this topic were:

- Large data sets vs large complex schemata
- Scalability is not only performance
- Visualizing large-schemas properly

Conclusions

The working group agreed that there is still a lot of things to do for convincing the industry about the practical applicability of current techniques for reasoning on the structural schema. Most of these things have been summarized along this section. However, the promising results achieved so far and the existence of several prototype tools that can be applied in practice allow us to be optimistic about the achievement of this ambitious goal. Having practical tools to show that all of this works was agreed to be a necessary condition for this purpose.

Participants:

- Achim D. Brucker (SAP Research – Karlsruhe, DE)
- Alessandro Artale (Free University of Bozen-Bolzano, IT)
- Alessandro Mosca (Free University of Bozen-Bolzano, IT)
- Bernhard Thalheim (Universität Kiel, DE)
- Carolina Dania (IMDEA Software Institute, ES)
- David W. Embley (Brigham Young University, US)
- Ernest Teniente (UPC – Barcelona, ES)
- Ingo Feinerer (TU Wien, AT)
- Mirco Kuhlmann (Universität Bremen, DE)
- Parke Godfrey (York University – Toronto, CA)
- Sophie Dupuy-Chessa (LIG – Grenoble, FR)
- Xavier Blanc (University of Bordeaux, FR)

4.2 Reasoning about the Conceptual Schema Components Capturing Dynamic Aspects

Diego Calvanese

The discussion in the working group started from the observation that the topic to be addressed is quite challenging for a variety of reasons. Indeed, there is a consensus about the key aspects that are of importance and need to be considered when modeling the structural aspects of a system and when reasoning over such a conceptualization. Instead, there was a consensus that when it comes to modeling and reasoning over the dynamic aspects of an information system, and hence its evolution over time, the situation is much less clear and not at all consolidated, both with respect to the properties to be modeled, and with respect to the formalisms to be adopted.

After a round in which each participant briefly introduced what it considered important aspects to be tackled in the discussion, the working group set up an ambitious agenda comprising the following list of points and questions that it intended to address, ordered by importance:

1. Which dynamic and/or temporal properties should be modeled? How should the structural and dynamic components be combined?
2. Which modeling formalisms, possibly based on logic, should be adopted for capturing the dynamic together with the static aspects of a system? In addition to the expressive power of the formalism, also the aspects related to the computational complexity and hence efficiency of reasoning should be taken into account, and hence discussed. Possibly, tractable fragments should be identified.
3. Identify specific problems, use cases, and scenarios related to dynamic aspects that come from industrial requirements (e.g., security).
4. Identify important design-time tasks where reasoning about dynamic aspects is of importance (e.g., exploratory design, analysis, planning, synthesis, verification).
5. Identify important run-time tasks where reasoning about dynamic aspects is of importance (e.g., analysis, validation, monitoring, mining).
6. Discuss the different levels of abstraction of models and their implementation.

When the group set out to discuss Item 1 of the above list, it became immediately clear that there was a very tight connection between the dynamic/temporal properties and the formalism to adopt for modeling them, so that Items 1 and 2 were actually discussed together. In fact, getting a clarification on these two points was considered almost a prerequisite for addressing further issues, so that the discussion on them took up almost the whole time available to the group. In the end, Items 3 and 6, although considered important, could not be addressed. Items 4 and 5 dealing respectively with the design-time and run-time tasks that could be supported by reasoning over dynamic aspects where considered and discussed together.

We summarize below the outcome of the discussions that took place in the group.

Properties and types of formalisms

There was an agreement that a convenient and general way to characterize the semantics of systems that evolve and change over time is by means of labeled transition systems over first-order structures (i.e., databases). Several dimensions for characterizing the formalisms that are able to specify and capture dynamic aspects were identified, specifically:

- adoption of a linear vs. a branching time model and formalism to specify dynamic properties;
- adoption of a propositional abstraction vs. first-order formalisms that are able to quantify over a single state of the (transition) system vs. first-order formalisms that are able to quantify across different states;
- how data should be incorporated: adoption of a pure first-order model vs. first-order model with built-in data types vs. first-order model enriched with additional structure, e.g., arithmetics;
- adoption of a discrete vs. a real-time model;
- adoption of a model based on a sequence/tree of states (i.e., databases) that are queried independently vs. a temporal database like model, where one can query arbitrarily the sequence/tree of states, also doing kind of complex joins across states;
- consideration or not of deontic and organizational aspects

Usability was considered an important dimensions that will have a strong impact on the choices of the formalism, but is an aspect that is orthogonal to all the ones mentioned above.

Concrete formalisms

The working group discussed also several concreted formalism that could be adopted for modeling and reasoning over dynamic aspects. Specifically, the following was considered and discussed:

- First-order mu-calculus was identified as a very general and powerful formalism for specifying dynamic properties.
- First-order variants of temporal logics, such as LTL, CTL, PDL, TLA+, B, and LDL were also mentioned; the availability of corresponding model checking tools (e.g., TLC) was considered an important aspect to take into account.
- Decidable fragments of the above languages should possibly be adopted. An example are temporal description logics, where to obtain decidability of reasoning over the combination of static and dynamic aspects (in some cases severe) restrictions on the expressive power of the modeling formalisms are made.
- Action-based languages that allow one to model processes, e.g., by means of rules were mentioned.
- It was observed that OCL, which combined with UML is already widely used to capture structural constraints of a system, can also be adopted to specify pre- and post-conditions of actions.

Interaction between the static and dynamic components

An important aspect was considered to be how the models and formalisms for the static and for the dynamic parts of the system interact with each other.

- One possibility is that the static and the dynamic aspects are part of the same conceptual model, which becomes a so-called temporal conceptual model.
- A different approach is to model the two aspects separately and then deal with their connection, which deals results in a structural conceptual model plus a business process.
- Also the form of the dynamic queries that can be posed over the evolution of the static component over time needs to be taken into account.

Conclusions

The working group agreed that there is a lack of a reference framework encompassing all relevant aspects. While such a reference framework on the one hand might be be too general in practice, on the other hand it could at least provide a way to connect the different approaches. It was also observed that when we consider models for the dynamics, there is even a lack of understanding and agreement of when a model/formalism should be called "conceptual", and what the right level of abstraction for a "dynamic conceptual model" actually is. Similarly to the case of modeling and reasoning over the structural aspects of a system, also in this case a key aspect is the trade-off between generality and expressive power on the one hand and decidability and complexity of inference on the other hand.

Participants:

- C. Marijke Keet (University of KwaZulu-Natal – Durban, ZA)
- Carsten Lutz (Universität Bremen, DE)
- Dániel Varró (Budapest Univ. of Technology & Economics, HU)
- Diego Calvanese (Free University Bozen-Bolzano, IT)
- Geri Georg (Colorado State University, US)
- Jerzy Marcinkowski (University of Wroclaw, PL)
- Jorge Lobo (IBM TJ Watson Research Center – Hawthorne, US)
- Marco A. Casanova (PUC – Rio de Janeiro, BR)
- Marco Montali (Free University Bozen-Bolzano, IT)
- Martin Gogolla (Universität Bremen, DE)
- Michael Leuschel (Heinrich-Heine-Universität Düsseldorf, DE)
- Michael Zakharyaschev (Birkbeck College London, GB)
- Xavier Oriol (UPC – Barcelona, ES)

4.3 New Challenges for Automated Reasoning on Conceptual Schemas

Sven Hartmann

Our working group started with a brainstorming session where everybody contributed ideas and opinions about emerging trends and challenges in our field of research. Soon the discussion focussed around the following set of questions:

- Which major trends are underway in database modelling and management? What will be the trends of tomorrow? Which trends will have a lasting impact?
- What challenges arise for the conceptual modelling of databases? What contributions can the conceptual modelling community make? What new research directions emerge for the conceptual modelling community?

- What support can automated reasoning provide to tackle these challenges? What concepts, theories, and methods must be developed by the automated reasoning community to enable solutions? What new research directions emerge for the automated reasoning community?

After a lively exchange of thoughts a range of trends and issues for research were identified as important and suggested for further discussion. These can be grouped into the following topic areas:

1. Big data and the exploration of conceptional schemas
2. Integration of database and software engineering formalisms and methods
3. Co-evolution of schemas and views
4. Updatable views
5. Assumptions of reasonings tasks
6. Meta modelling
7. Research integration
8. Benchmarks
9. Quality assurances and cost models

During the discussion in our group it soon became apparant that some of the upcoming research issues are extensions of established problems that have motivated research on automated reasoning for conceptual schemas in the past and still await answers that unlock them for uptake by database practinioneers.

Big Data and Schema Exploration

Big Data is one of today's mega trends that leads to a multitude of new research issues. The group shared opinions on whether and how conceptual modelling can help to understand and process massive datasets of heterogenous data. The following items were suggested for further investigation:

- Can we use automated reasoning techniques to extract / adjust / enhance conceptual information from big data (such as sensor data or stream data)?
- How would that be different from data mining?
- How can existing conceptual models be combined with data mining techniques to improve concept models?
- What formalisms are needed to capture conceptual information in big data.

Database vs software development

In view of the increasing complexity of future information systems the group found it necessary to develop and deploy research-led strategies for integrating formalisms, methods and practices used in database and software development.

- How can constraint checking efforts in DBMS and application software be integrated?
- What guarantees do software developers need about databases?
- Does correctness / completeness really matter?
- How can these guarantees be best communicated?
- Do we need to map them to implementation code?
- How to establish a global perspective on constraint enforcement?

Co-evolution of Schemas and Views

The group discussed challenges that arise from changing information needs in complex information systems. There was also agreement that conceptual modelling in practice has diverse stakeholders that require tailored views on databases during the entire life cycle.

- How can conceptual changes of views be propagated to the underlying schema?
- What updates can be supported?
- How can schema/view mappings best evolved?
- What constraints can be supported?
- What criteria need to be considered?
- How can automated reasoning help?
- Can we understand and contribute to tools like Hibernate?

Updatable views

The discussion addressed research challenges for automated reasoning that arise when databases are updated by diverse users and application programs through tailored views.

- How can data updates on views be propagated to the underlying conceptual schema?
- What updates can be supported?
- What constraints can be supported?
- What criteria need to be considered? (lossless, inference-proof, . . .)
- How can automated reasoning help?

Assumptions of reasonings tasks

The group discussed the discrepancy between the challenges that database designers face in practice and the answers that the research community is able to provide, and asked for reasons.

- How to bridge the mismatch between the relational model and SQL?
- Can our methods handle partial information?
- Can our methods handle duplicates?
- How can SQL features be handled in the logic languages that we use?

Meta modelling

In view of the diversity of conceptual modelling frameworks in use the group agreed that automated reasoning can only be successful if the underlying semantics is revealed and formalized. The opportunities of model-based approaches and schema translations were discussed, too.

- Do we need semantics for meta models?
- Can we find and justify patterns for model transformations in conceptual modelling?
- Are purely syntactical transformations desirable / achievable?
- How can graphical and textual languages for schema declaration be used simultaneously to provide optimal support for designers?

Research integration

There was agreement that the tackling emerging research challenges would benefit from joint efforts of the community which requires a common understanding and extended collaboration.

- Do we understand each other? Even if we know about different terminologies, are we always aware of the one we are using?
- We use different methods for constraint handling, different languages for declaring constraints, different constraint classes – how can these efforts be integrated?
- Do we need new approaches to tackle the conceptual complexity of emerging applications?
- Can modelling and reasoning be combined with simulation?

Benchmarks

We discussed what ingredients would be most helpful to facilitate future collaboration in research, and identified commonly accepted benchmarks as one important enabling tool.

- Are there proper benchmarks for conceptual modelling tasks?
- What qualifies them as benchmarks?
- How to make test cases transparent and applicable by the community?

Quality assurances and cost models

Members of our group asked what obstacles hamper the uptake of new research achievements into practice, and what is needed to overcome them. Quality assurances in the vein of service-level agreements were considered as one promising approach and suggested for investigation.

- Is it necessary / possible / desirable to predict the impacts of design decisions at conceptual level on the implementation level?
- What kind of cost models do we need / want?
- What assurances can be provided to database users?
- Is it enough to say “we do it best possible” or do we need quantitative statements?

Participants:

- Elena V. Ravve (ORT Braude College – Karmiel, IL)
- Enrico Franconi (Free University Bozen-Bolzano, IT)
- Guillem Rull (UPC – Barcelona, ES)
- Joachim Biskup (TU Dortmund, DE)
- Klaus-Dieter Schewe (Software Competence Center – Hagenberg, AT)
- Mira Balaban (Ben Gurion University – Beer Sheva, IL)
- Qing Wang (Australian National University, AU)
- Roman Kontchakov (Birkbeck College London, GB)
- Stephan Mäs (TU Dresden, DE)
- Stephen J. Hegner (University of Umeå, SE)
- Sven Hartmann (TU Clausthal, DE)
- Thomas Baar (Hochschule für Technik und Wirtschaft – Berlin, DE)

5 Seminar program

The program of the different sessions is given below.

Session 1: Reasoning on the Structural Schema (I)

- UML class diagrams – decision, identification and repair of correctness and quality problems, *Mira Balaban*
- OCL2FOL: Using SMT solvers to automatically reason on conceptual schemata with OCL constraints, *Carolina Dania*
- Reasoning techniques for conceptual models, *Alessandro Artale*
- Toward an ontology-driven unifying metamodel for UML Class Diagrams, *Maria Keet*

Session 2: Reasoning on the Structural Schema (II)

- "Automating reasoning on conceptual schemas" in FamilySearch—a large-scale reasoning application, *David W. Embley*
- Incremental inconsistencies detection with low memory overhead, *Xavier Blanc*
- Constraints on class diagrams, *Ingo Feinerer*
- At SAP, class models are rarely used as they are "too close to real code", *Achim D. Brucker*

Session 3: Reasoning on the Structural Schema (III)

- Reasoning in ORM, *Enrico Franconi*
- Exploring UML and OCL Model Properties with Relational Logic, *Martin Gogolla*
- AuRUS: Automated reasoning on UML schemas, *Ernest Teniente*
- ProB: Solving constraints on large data and higher-order formal models, *Michael Leuschel*

Session 4: Extensions

- Preliminary report on an algebra of lightweight ontologies, *Marco A. Casanova*
- Temporal extended conceptual models, *Roman Kontchakov*
- Reasoning on conceptual schemas of spatial data, *Stephan Mäs*
- Validation of complex domain-specific modeling languages, *Dániel Varró*

Session 5: Reasoning about the Dynamics

- View design for updates, *Stephen Hegner*
- Reasoning about the effect of structural events in UML conceptual schemas, *Xavier Oriol*
- Automated reasoning for security and compliance properties of business processes, *Achim D. Brucker*
- Unified approaches for modeling and reasoning over processes and data, *Diego Calvanese*

Session 6: New Challenges

- The curse of restructuring in dependency theory, *Klaus-Dieter Schewe*
- A declarative approach to distributed computing, *Jorge Lobo*
- On BDD, finite controllability and the BDD/FC conjecture, *Jerzy Marcinkowski*
- Metrics for visual notations, *Sophie Dupuy-Chessa*

Session 7: Reasoning about Mappings

- Reasoning about dependencies in schema mappings, *Qing Wang*
- Semantic-based mappings, *Guillem Rull*
- Relationship between approaches to ontology-based data access and object relational techniques, *Marco Montali*
- Armstrong instances as an aid for automated reasoning, *Sven Hartmann*

Session 8: Reasoning about Dependencies

- Information and dependency preserving BCNF decomposition algorithm via attribute splitting, *Elena V. Ravve*
- Visual reasoning with (functional) dependencies, *Bernhard Thalheim*
- Representation of instance-derivations based on dependencies, *Joachim Biskup*
- Reasoning over order dependencies for relational schema, *Parke Godfrey*

Participants

- Alessandro Artale
  Free Univ. of Bozen-Bolzano, IT
- Thomas Baar
  Hochschule für Technik und Wirtschaft – Berlin, DE
- Mira Balaban
  Ben Gurion University – Beer Sheva, IL
- Joachim Biskup
  TU Dortmund, DE
- Xavier Blanc
  University of Bordeaux, FR
- Achim D. Brucker
  SAP Research – Karlsruhe, DE
- Diego Calvanese
  Free Univ. of Bozen-Bolzano, IT
- Marco A. Casanova
  PUC – Rio de Janeiro, BR
- Carolina Dania
  IMDEA Software Institute, ES
- Sophie Dupuy-Chessa
  LIG – Grenoble, FR
- David W. Embley
  Brigham Young University, US
- Ingo Feinerer
  TU Wien, AT
- Enrico Franconi
  Free Univ. of Bozen-Bolzano, IT
- Geri Georg
  Colorado State University, US
- Parke Godfrey
  York University – Toronto, CA
- Martin Gogolla
  Universität Bremen, DE
- Sven Hartmann
  TU Clausthal, DE
- Stephen J. Hegner
  University of Umeå, SE
- C. Maria Keet
  University of KwaZulu-Natal – Durban, ZA
- Roman Kontchakov
  Birkbeck College London, GB
- Mirco Kuhlmann
  Universität Bremen, DE
- Michael Leuschel
  Heinrich-Heine-Universität Düsseldorf, DE
- Jorge Lobo
  Universitat Pompeu Fabra – Barcelona, ES
- Carsten Lutz
  Universität Bremen, DE
- Stephan Mäs
  TU Dresden, DE
- Jerzy Marcinkowski
  University of Wroclaw, PL
- Marco Montali
  Free Univ. of Bozen-Bolzano, IT
- Alessandro Mosca
  Free Univ. of Bozen-Bolzano, IT
- Xavier Oriol
  UPC – Barcelona, ES
- Elena V. Ravve
  ORT Braude College – Karmiel, IL
- Guillem Rull
  UPC – Barcelona, ES
- Klaus-Dieter Schewe
  Software Competence Center – Hagenberg, AT
- Ernest Teniente
  UPC – Barcelona, ES
- Bernhard Thalheim
  Universität Kiel, DE
- Dániel Varró
  Budapest Univ. of Technology & Economics, HU
- Qing Wang
  Australian National Univ., AU
- Michael Zakharyaschev
  Birkbeck College London, GB

Report from Dagstuhl Seminar 13212

Computational Methods Aiding Early-Stage Drug Design

Edited by

Andreas Bender[1], Hinrich Göhlmann[2], Sepp Hochreiter[3], and Ziv Shkedy[4]

1 **University of Cambridge, GB,** ab454@cam.ac.uk
2 **Janssen Pharmaceuticals – Beerse, BE**
3 **University of Linz, AT,** hochreit@bioinf.jku.at
4 **Hasselt University – Diepenbeek, BE,** ziv.shkedy@uhasselt.be

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 13212 “Computational Methods Aiding Early-Stage Drug Design”. The aim of the seminar was to bring scientists working on various aspects of drug discovery, genomic technologies and computational science (e.g., bioinformatics, chemoinformatics, machine learning, and statistics) together to explore how high dimensional data sets created by genomic technologies can be integrated to identify functional manifestations of drug actions on living cells early in the drug discovery process.

Seminar 19.–24. May, 2013 – www.dagstuhl.de/13212
1998 ACM Subject Classification G.3 Probability and Statistics, I.5 Pattern Recognition, J.2 Physical Sciences and Engineering, J.3 Life and Medical Sciences
Keywords and phrases Bioinformatics, Chemoinformatics, Machine learning, Statistics, Interdisciplinary applications
Digital Object Identifier 10.4230/DagRep.3.5.78

1 Executive Summary

Andreas Bender
Hinrich Göhlmann
Sepp Hochreiter
Ziv Shkedy

Besides discussing scientific findings enabled by computational approaches, the seminar successfully stimulated discussions between scientists from different disciplines and provided an exceptional opportunity to create mutual understanding of the various challenges and opportunities. It created understanding for technical terms and concepts and served as a catalyst to explore new ideas.

As a concrete example, it challenged the feasibility of utilizing chemical structure information for identifying correlations with biological data. Rather than attempting to define a most suitable way of translating chemical structure information into computer understandable form (e.g., via fingerprinting algorithms such as ECFP), the notion of utilizing functional readouts such as gene expression profiles was favored for prioritizing candidate drugs that demonstrate a favorable balance of desired and undesired compound effects.

Computational Methods Aiding Early-Stage Drug Design, *Dagstuhl Reports*, Vol. 3, Issue 5, pp. 78–94
Editors: Andreas Bender, Hinrich Göhlmann, Sepp Hochreiter, and Ziv Shkedy
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Table of Contents

Executive Summary

Overview of Talks

Scientific Background

3 Overview of Talks

3.1 Enriched methods for analysis of high-dimensional data

Dhammika Amaratunga (Johnson & Johnson, US)

High-dimensional data are characterized as having an enormous number of features and relatively few samples. Exploration and classification of such data is an important aspect of bioinformatics and cheminformatics work. One of the challenges presented by such situations is that only a very small percentage of the features actually carries classification information; the other features add noise or carry secondary signals that could seriously obscure the true signal. In such situations, it is helpful to use methods that highlight features of the data that are most likely to be informative. We refer to such methods as enriched methods. Enriched methods have been developed for single-run procedures, such as SVD, as well for multiple-run (ensemble) procedures, such as Random Forest. Here we will discuss many issues that arise in this context and demonstrate the value of enriched methods in analyzing high-dimensional data.

3.2 Similarity-Based Clustering of Compounds and its Application to Knowledge Discovery from Kernel-based QSAR Models

Ulrich Bodenhofer (University of Linz, AT)

Joint work of Bodenhofer, Ulrich; Klambauer, Günter; Palme, Johannes; Kothmeier, Andreas; Hochreiter, Sepp
Main reference U. Bodenhofer, A. Kothmeier, S. Hochreiter, "APCluster: an R package for affinity propagation clustering," Bioinformatics, 27:2463–2464, 2011.
URL http://dx.doi.org/10.1093/bioinformatics/btr406

Quantitative structure-activity relationships (QSAR) have become a standard methodology in computational pharmacology and computer-aided drug design. In recent years, kernel-based approaches have been established as an alternative to traditional feature-based approaches using chemical descriptors or structural descriptors like ECFP fingerprints. Kernels can incorporate virtually any kind of chemical or structural information, as far as 3D structures. Many common kernels even facilitate good model interpretability similarly to feature- based approaches. This can be accomplished by the extraction of explicit feature weights and the superimposition of feature weights on given chemical structures to highlight sub-structures that are particulary relevant for the given modeling task. The only painful drawback of kernel approaches is their poor ability to scale to larger data sets.

In this contribution, we advocate the use of affinity propagation (AP) clustering for selecting representative samples/compounds, the so-called exemplars, with the following two possible applications: (1) If a set of exemplars is available, the kernel matrix can be compacted by removing all columns that do not correspond to exemplars. The Potential Support Vector Machine (P-SVM), for example, can process such non-quadratic kernel matrices and, thereby, leverage the scaling problem mentioned above. (2) As mentioned above, the superimposition of feature weights on chemical structures to highlight relevant sub-structures is an excellent tool for knowledge acquisition, but it can only be applied to a very limited number of samples. It seems natural to priorize samples/compounds that are known — on the basis of an objective procedure — to be most representative for the entire data set of compounds.

It is worth to note that the use of AP clustering is essential for these two applications. Firstly, AP is similarity-based, therefore, it is not limited to explicit feature representations, but can be used for all kinds of kernels as well. Secondly, AP is able to compute exemplars that are members of the original data sets (as opposed to hypothetical averages used by k-means clustering). Last but not least, AP is efficient and, with appropriate computational strategies, it scales to large data sets.

3.3 Protein family focused, structure enabled chemical probes, to accelerate the discovery of new targets

Chas Bountra (University of Oxford – Nuffield College, GB)

The discovery of "pioneer medicines" (i.e. those acting via novel molecular targets) has proven to be an immensely complex, long term, expensive and high risk endeavour. During my presentation, I will discuss

- our focus on novel human epigenetic protein families,
- the generation of freely available inhibitors, and
- the partnership with many academic and industrial labs

I will describe progress with novel inhibitors for bromodomain and demethylase proteins, and their use in identifying new targets in can cer, inflammatory and neuro-psychiatric diseases.

3.4 Combining transcriptomics, bioassays and chemistry to aid drug discovery

Hinrich Göhlmann (Janssen Pharmaceutica – Beerse, BE)

Joint work of QSTAR Consortium

The joint talk introduced the participants of the seminar to the steps and challenges of the drug discovery and development process. Janssen has collaborated over the past years with several universities to explore how transcriptomic data can be used to aid and overcome these challenges. With the financial support of the flemish IWT we have initially developed algorithms and approaches for defining and prioritizing compound clusters of equipotent compounds that were active in a phenotypic screen. In the second research and development project (QSTAR) we are attempting to combine transcriptomic data with data of bioassays and chemical structure information. Modelling either two of the three data types or jointly modelling all data we investigate what correlations are present in the data. As an essential tool for facilitating the collaboration between the different partners we have developed data processing pipelines that generate robust data structures that link all three data types via unified compound identifiers (InChIKey). By jointly creating R packages for data access as well as data analysis we hope to create the foundation that will allow us and other interested research teams to investigate what connections are present in the data and how we can use the information to aid drug discovery in the future.

3.5 False discovery proportions of gene lists prioritized by the user

Jelle J. Goeman (Leiden University Medical Center, NL)

Joint work of Goeman, Jelle J.; Solari, Aldo
Main reference J. J. Goeman, A. Solari, "Multiple Testing for Exploratory Research," Statistical Science, 26(4):584–597, 2011.
URL http://dx.doi.org/doi:10.1214/11-STS356

Motivated by the practice of exploratory research, we formulate an approach to multiple testing that reverses the conventional roles of the user and the multiple testing procedure. Traditionally, the user chooses the error criterion, and the procedure the resulting rejected set. Instead, we propose to let the user choose the rejected set freely, and to let the multiple testing procedure return a confidence statement on the number of false rejections incurred. In our approach, such confidence statements are simultaneous for all choices of the rejected set, so that post hoc selection of the rejected set does not compromise their validity. The proposed reversal of roles requires nothing more than a review of the familiar closed testing procedure, but with a focus on the non-consonant rejections that this procedure makes. We suggest several shortcuts to avoid the computational problems associated with closed testing.

3.6 Systematic mapping of synthetic genetic interactions with combinatorial RNAi

Wolfgang Huber (EMBL Heidelberg, DE)

Biological systems are able to buffer the effects of individual mutations, and disease outcomes often depend on the combination of multiple genetic variants. Genetic interactions have been systematically measured in yeast and enabled the placement of genes into functional modules and the delineation of networks between modules at unprecedented coverage. However, such approaches have not been feasible for higher organisms. In this talk, I will report on our high-resolution genetic interaction maps of chromatin-related genes in Drosophila and human cells, obtained by combinatorial perturbation via RNA interference and single-cell phenotyping by automated imaging. Genetic interaction profiles were obtained by measuring multiple, non-redundant cellular phenotypes. The analysis of profiles revealed functional modules, among them many conserved protein complexes. Comparison with yeast showed a consistent, evolutionarily conserved pattern of genetic interactions for the substructures of the mediator complex, but also revealed the functional divergence of the kinase module Cdk8 and CycC in Drosophila. Genetic epistasis is an unresolved frontier of cancer genetics, as sequencing projects found that combinations of multiple, partially alternative and individually rare mutations lead to equivalent phenotypes. To dissect such interdependencies, we mapped recently reported recurrent cancer mutations onto our network and grouped them into clusters of putatively equivalent network functions.

3.7 Drug-induced transcriptional modules in mammalian biology: implications for drug repositioning and resistance

Murat Iskar (EMBL Heidelberg, DE)

Joint work of Iskar, Murat; Zeller, Georg; Blattmann, Peter; Campillos, Monica; Kuhn, Michael; Kaminska, Katarzyna; Runz, Heiko; Gavin, Anne-Claude; Pepperkok, Rainer; van Noort, Vera; Bork, Peer
Main reference M. Iskar, G. Zeller, P. Blattmann, M. Campillos, M. Kuhn, K.H. Kaminska, H. Runz, A.-C. Gavin, R. Pepperkok, V. van Noort, P. Bork, "Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding," Molecular Systems Biology, 9:662, 2013.
URL http://dx.doi.org/10.1038/msb.2013.20

In recent years, the publicly available data on small molecules has increased dramatically. Integrative analysis of these heterogeneous resources enables us to gain a better understanding of drug action in biological systems. Genome-wide expression profiling of cells treated with drugs summarizes the pharmacological and toxicological effects of these perturbations at the molecular level and further help us to bridge between the molecular basis of drug action and their phenotypic consequences. To systematically explore the biological responses of mammalian cells to a diverse set of chemical perturbations, we generated a comprehensive collection of drug-induced transcriptional modules from existing microarray data on drug-treated human cell lines and rat liver. More than 70% of these modules were identified in multiple human cell lines and 15% were conserved across organisms of human and rat, representing a lower limit. We systematically characterized these modules and could link antipsychotic drugs to sterol and cholesterol biosynthesis, providing an explanation for the metabolic side effects reported for these drugs. Moreover, we could identify novel functional roles for hypothetical genes, e.g. ten new modulators of cellular cholesterol levels and novel therapeutic roles for several drugs, e.g. new cell cycle blockers and modulators of alpha-adrenergic, PPAR and estrogen receptors. Our work not only quantifies the conservation of transcriptional responses across biological systems, but also identifies novel associations between drug-induced transcriptional modules, drug targets and side effects.

3.8 Semi-supervised investigation of association of gene expression with structural fingerprints of chemical compounds

Adetayo Kasim (Durham University, GB)

Main reference Y. Li, K. Tu, S. Zheng, J. Wang, Y. Li, P. Hao, X. Li, "Association of feature gene expression with structural fingerprints of chemical compounds," Journal of Bioinformatics and Computational Biology, 9(4): 503–519, 2011.
URL http://dx.doi.org/10.1142/S0219720011005446

Exploring the relationship between a chemical structure and its biological function is of great importance for drug discovery. Whilst many studies attempt to introduce transcriptomics data into chemical function, little effort has been made to link structural fingerprints of compounds with defined intracellular functions such as target related pathways or expression of particular set of genes. Li *et al.* (2011) propose an approach to associate structural differences between compounds with the expression level of a defined set of genes by performing clustering on chemical structures to find differentially expressed genes between adjacent clusters of compounds from the same node. The identified set of genes were further subjected to compounds re-classification to evaluate the accuracy of the prediction based on gene expression and chemical structures.

We propose a semi-supervised approach for investigation of association of gene expression with structural finger prints. Our approach starts with unsupervised biclustering of gene expression to identify subset of genes and compound with target related pathways. The expression levels of the relevant genes are used to weight structural fingerprints of chemical compounds to obtain clusters of compounds with a common set of structural fingerprints and similar level of gene expression. A similar approach was also applied to identify clusters of compounds with a common set of fingerprints and similar bioassay level.

3.9 Multi-view learning for drug sensitivity prediction

Samuel Kaski (Aalto University, FI)

We are developing machine learning methods for integrating multiple high- dimensional data sources. In the unsupervised task of decomposing the sources into shared and source-specific components, the new Bayesian Canonical Correlation Analyses and Group Factor Analyses can be applied to study omics- wide effects of chemical structures. The supervised personalized medicine task of predicting drug sensitivity based on multiple genomic measurement sources, can be addressed by a combination of multi-view and multi-task learning. This is joint work with several people from my group, and collaborators from Institute for Molecular Medicine Finland FIMM. For more details and code see http://research.ics.aalto.fi/mi.

3.10 Detecting differentially expressed genes in RNA-Seq drug design studies

Guenter Klambauer (University of Linz, AT)

Joint work of Klambauer, Guenter; Unterhiner, Thomas; Hochreiter, Sepp
URL http://www.bioinf.jku.at/software/dexus/

Detection of differential expression in RNA-Seq data is currently limited to studies in which two or more sample conditions are known a priori. However, these biological conditions are typically unknown in drug design studies. We present DEXUS for detecting differential expression in RNA-Seq data for which the sample conditions are unknown. DEXUS models read counts as a finite mixture of negative binomial distributions in which each mixture component corresponds to a condition. A transcript is considered differentially expressed if modeling of its read counts requires more than one condition. DEXUS decomposes read count variation into variation due to noise and variation due to differential expression. Evidence of differential expression is measured by the informative/non-informative (I/NI) value, which allows differentially expressed transcripts to be extracted at a desired specificity (significance level) or sensitivity (power). DEXUS performed excellently in identifying differentially expressed transcripts in data with unknown conditions. On 2,400 simulated data sets, I/NI value thresholds of 0.025, 0.05, and 0.1 yielded average specificities of 92%, 97%, and 99%

at sensitivities of 76%, 61%, and 38% respectively. On real-world data sets, DEXUS was able to detect differentially expressed transcripts related to sex, species, tissue, structural variants, or eQTLs.

3.11 Intestinal microbiota, individuality and health

Leo Lahti (Wageningen University, NL)

Joint work of Lahti, Leo; Jarkko Salojärvi; Anne Salonen; Marten Scheffer; Willem M de Vos
URL http://microbiome.github.com

Diverse microbial communities inhabit the human gastrointestinal tract, where hundreds of distinct bacterial phylotypes and a trillion bacterial cells per gram in a healthy adult individual can be encountered. This ecosystem constitutes a virtual metabolic organ that has a central role in nutrition, immune system and other bodily functions, and a profound impact on our well-being.

Recent accumulation of high-throughput profiling data sets is now for the first time enabling global characterization of the overall composition and variability of this intestinal microbiota. Integration of phylogenetic profiling data of a thousand phylotypes across thousands of human individuals scales up the current analyses by an order of magnitude based on the Human Intestinal Tract chip (HITChip), a phylogenetic microarray has enabled standardized data collection of over one thousand gut-specific bacterial phylotypes including many less abundant species that cannot be cultivated in a laboratory and whose functional role is less well known.

The analysis reveals huge inter-individual variability in microbial diversity as well as alternative ecosystem states that are associated with personal environmental and phenotypic factors such as ageing, overweight, host metabolism, and health status. We will discuss how these recent observations provide new insights into the role of our co-evolved microbial partners in individual health and well-being, as well as guidance for the design and interpretation of future studies.

3.12 Library-Scale Gene-Expression Profiling and Digital Open Innovation

Justin Lamb (Genometry Inc – Cambridge, US)

The Broad Institute's Connectivity Map project (www.broadinstitute.org/cmap) has demonstrated the value of a database of gene-expression profiles derived from cultured cells treated with a large collection of bioactive small molecules for drug discovery and development applications. It has also shown how exposing these data and allied search algorithms to the global biomedical-research community through a simple self-service webtool can successfully digitize and democratize the small-molecule screening process. The talk will review this earlier work then describe our efforts to greatly expand the Connectivity Map using a novel high-throughput low-cost gene-expression profiling technology we have developed. The idea that a comparable system populated with expression profiles of a pharmaceutical company's proprietary chemical matter could serve as an efficient open-innovation platform will also be discussed.

3.13 A Maximum Common Subgraph Kernel Method for Predicting the Chromosome Aberration Test

Johannes Mohr (TU Berlin, DE)

Joint work of Mohr, Johannes; Jain, B. ; Sutter, A.; Ter Laak, A.; Steger-Hartmann, T.; Heinrich, H.; Obermayer, K.
Main reference J. Mohr, B. Jain, A. Sutter, A. Ter Laak, T. Steger-Hartmann, N. Heinrich, K. Obermayer, "Maximum Common Subgraph Kernel Method for Predicting the Chromosome Aberration Test," J. Chem. Inf. Modeling, 50(10), pp. 1821–1838, 2010.
URL http://dx.doi.org/10.1021/ci900367j

The chromosome aberration test is frequently used for the assessment of the potential of chemicals and drugs to elicit genetic damage in mammalian cells in vitro. Due to the limitations of experimental genotoxicity testing in early drug discovery phases, a model to predict the chromosome aberration test yielding high accuracy and providing guidance for structure optimization is urgently needed. In this talk I will present a machine learning approach for predicting the outcome of this assay based on the structure of the investigated compound. It combines a maximum common subgraph kernel for measuring the similarity of two chemical graphs with the potential support vector machine for classification. The approach allows visualizing structural elements with high positive or negative contribution to the class decision.

3.14 Recursive Neural Networks for Undirected Graphs and Neural Network Pairwise Interaction Fields for annotating 2D and 3D small molecules

Gianluca Pollastri (University College Dublin, IE)

Joint work of Pollastri, Gianluca; Lusci, Alessandro; Baldi, Pierre
Main reference A.Lusci, G.Pollastri, P. Baldi, "Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules," Journal of Chemical Information and Modeling, 53(7), pp. 1563–1575, 2013
URL http://pubs.acs.org/doi/abs/10.1021/ci400187y

I introduced two ways of "wiring" Artificial Neural Networks, that we have developed in my group, which can deal with structured data in the form of graphs. The first one, UG-RNN (Recursive Neural Networks for Undirected Graphs) factorises an undirected graph into a number of directed graphs that are used to transfer contextual information, and compresses this contextual information into a feature vector which can be mapped into a desired target property. The second model, NN-PIF (Neural Network Pairwise Interaction Field), subdivides a graph into all the pairwise interactions between its nodes (each, potentially, represented alongside a context of neighbours) and maps these pairwise interactions into a feature vector, which is then mapped into a target property. In both cases the feature vector is automatically generated, and is effectively a fixed-size property-driven adaptive representation of the input. We have tested both models on a number of problems in the space of small molecules, in their 2D representation (in the case of UG-RNN) and 3D representation (for NN-PIF). I described the results we have obtained in these tests, which are generally comparable, and often superior, to those obtained by state of the art 2D and 3D kernels on the same sets. I also speculated on the feature vectors produced by both models, and on how they may be used for mapping the input space.

3.15 The nature of gene signature development

Willem Talloen (Janssen Pharmaceutica – Beerse, BE)

We now have entered the post-genomic era with much hope to harvest some of the fruits hidden in the genomic text. At the same time, the current difficulties faced by pharma research to discover generally applicable block-buster drugs have lead to think in terms of personalized medicine. Consequently, high hopes are on clinical opportunities for gene expression-based prediction of illness or drug response discovered using high-content technologies such as microarrays or RNAseq.

The 'omics revolution was also warmly welcomed by data analysts as its data properties imposed new and interesting statistical challenges. For example, the quest for biomarkers in the context of personalized medicine has made many statisticians and bioinformaticians think about classification models that are robust against overfitting for generation of molecular signatures.

Here, we will challenge that this enthusiasm made many researchers forget to think about the practical applicability and the biological nature, and hence clinical relevance of these developed classification algorithms. For example, the rationale behind signatures consisting out of many genes is generally overlooked. How these genes should be aggregated into one composite index (i.e., the marker) so as to reflect the underlying biology as well as to remain generalizable will be discussed in this presentation.

3.16 Scaling bioinformatics algorithms

Oswaldo Trelles (University of Malaga, ES)

Large scale genomics projects exploiting high throughput leading technology have produced and continue to produce massive data sets with exponential growing rates. So far, only a small part of this data can be abstracted, managed and processed, giving an incomplete understanding of the biological process being observed. The lack of processing power is a bottle neck in acquiring results.

A promising approach to address the processing of such massive data sets is the creation of new computer software that makes effective use of parallel and cloud computing.

Comparative genomics is a good example since it includes all the ingredients: huge and ever growing datasets, complex applications that demands large computational resources and new mathematical and statistical models for analysing and synthetizing genomic information.

This talk will provide an overview of cloud computing -from the user perspective- and the ways to exploit it with a real implementation in the framework of the Mr.SYmbiomath project

3.17 Minor Variant Detection In Virology with Model Based Clustering

Bie Verbist (Ghent University, BE)

Deep-sequencing is one of the applications of the new massively parallel sequencing (MPS) technologies allowing for an in-depth characterization of sequence variation in more complex populations, including low-frequency viral strains. However, MPS technology-associated errors in the resulting DNA sequences may occur up to equal or even higher frequency than the truly present mutations in the biological sample, impeding a powerful assessment of low- frequency virus mutations. As there are no obvious solutions to reduce the technical noise by further improvements of the technology platform, we believe that the search for statistical algorithms that can better correct the technical noise can be pivotal. Therefore algorithms that increase detection power in presence of technical noise and quantify base-call reliability are required. Phred-like quality scores, provided with the base-calls are such a quantification of the base-call reliability. These quality scores together with other covariates determine the multinomial model structure in a model-based clustering approach which will allow identification of viral quasi-species. This research program was granted by IWT, a governmental agency for Innovation by Science and Technology in Flanders, Belgium.

4 Scientific Background

(Written by Andreas Bender, Hinrich Göhlmann, Sepp Hochreiter, Ziv Shkedy)

4.1 Motivation

The efficiency and effectiveness of drug discovery has been challenged over the past years as increasing numbers of drug candidates failed to reach the market and patients. Accordingly, many efforts are underway to increase the productivity of the R&D process and avoid expensive late-stage clinical failures. A key concept is to de-risk drug candidates during the early preclinical stages. Toward this end, it is essential to reduce the time gap between the selection of promising candidate compounds (chemotypes) and the identification of potential side effects in later toxicity studies. In other words, relevant biologically data on the various desired and undesired effects of compounds need to be acquired early on in the research process.

At the same time various modern molecular biology technologies (e.g., next-generation sequencing, microarrays, high content screening) have advanced our understanding of the molecular basis of diseases and drug actions. One approach to increase the productivity of drug discovery is to complement traditional pharmacology approaches by using modern molecular biology technologies together with computational techniques, e.g., studying the effects of drugs on a cell line. These new opportunities, in particular, the integration of gene expression data on a large scale, complement the established methods in computational chemistry that are focused on the chemical structures. Consequently, drug designers have to become able to interrelate and interpret transciptomic, genetic, proteomic, metabolomic, and assay data.

The aim of the seminar "Computational Methods Aiding Early-Stage Drug Design" was to bring scientists working on various aspects of drug discovery, genomic technologies and computational science (e.g., bioinformatics, chemoinformatics, machine learning, and statistics) together to explore how high dimensional data sets created by genomic technologies can be integrated to identify functional manifestations of drug actions on living cells early in the drug discovery process.

4.2 Previous Work and State-of-the-Art

The focus of the seminar is on utilizing computational methods in early-stage drug design. Pharmaceutical companies have large libraries of compounds at their disposal which they investigate for their potential medical applicability. In the early stage, compounds that seem promising to become drugs are selected for subsequent drug development phases.

Currently, two main branches to drug design are employed. On the one hand, structure-based drug design is based on predicting which compound binds to a specific target under investigation. Standard methods for target-compound docking are only feasible for screening a small set of compounds and include docking algorithms, molecular dynamics (force fields), and even simulations of quantum mechanics effects, which are all based on the 3D structure of the target. For screening many compounds simultaneously, specific features of the 3D structure of the target are incorporated into computational methods that predict whether a compound binds to the target.

On the other hand, the current practice in many pharmaceutical companies is to apply ligand-based drug design, i.e. to screen their huge compound libraries for possible drug candidates. Ligand-based drug design is often based on "analoging", where the activity of compounds that are similar to known active compounds is predicted by means of computational models. These analoging models are generated on the basis of similarity of compounds whose biological activity or target specificity has been verified experimentally. Analoging is mostly based on pharmacophores (target-specific functional groups in the compounds) and structure-activity relationship (QSAR) models which are used to represent the relationship between compound properties and the biological activity of the compound.

Ligand-based drug design typically consists of the following steps (see Fig. 1):

1. **Compound filtering**: From a large library of chemical compounds, a sub-selection of feasible compounds is chosen. Criteria for this filtering include, but need not be limited to, the following: solubility, permeability, (non-)toxicity, chemical feasibility, costs, and patent issues.
2. **Compound selection**: From the set of feasible compounds, those are chosen that seem to have the desired effect related to the disease under investigation. This selection may be based on phenotypic screens, disease-specific bioassays, or other techniques. Analoging is an approach to identify new active (having the desired effect) candidates, which are found by their similarity to existing active compounds.

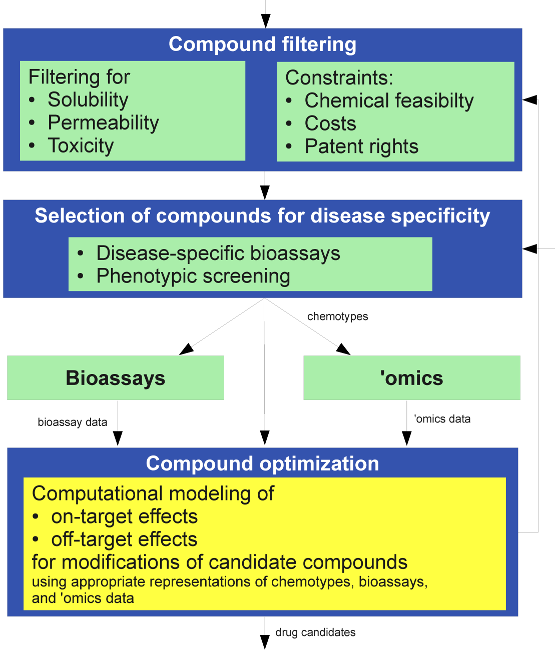


Figure 1 Overview of the steps during ligand-based drug design.

3. **Bioassay measurements**: A large set of bioassays and various ’omics technologies (e.g. gene expression measurements) are used to assess on-target and off-target effects of the selected compounds.
4. **Compound optimization**: a small set of so-called lead compounds is selected; potentially promising modifications of these lead compounds are considered in order to maximize target effects and minimize off-target effects. The most promising lead compounds are fed back into steps 1, 2, or 3 to verify their expected properties experimentally.

That only a limited number of modifications can be studied experimentally is the central bottleneck of this procedure. Computational methods provide a powerful tool set to avoid this bottleneck. They are predominantly applied in the design-make-test cycle for compound optimization to screen large sets of compound modifications for the most promising candidates (analoging) that can then be tested experimentally. Secondly, computational methods are also highly relevant for predicting the solubility, permeability and toxicity of the compounds which further reduces the number of compounds which need to be tested experimentally.

For compound optimization, ideally, three kinds of data sets are available for each compound (see Fig. 2):

1. **chemical structure and properties**
2. **’omics data**, e.g. transcriptome (in one condition/cell line or in several conditions/cell lines)
3. **phenotypic data** (biological assays)

Using these data, compounds are optimized for being more effective (on-target) and, at the same time, for having less side effects (off-target). On-target and off-target effects are determined by bioassays and by ’omics technologies.

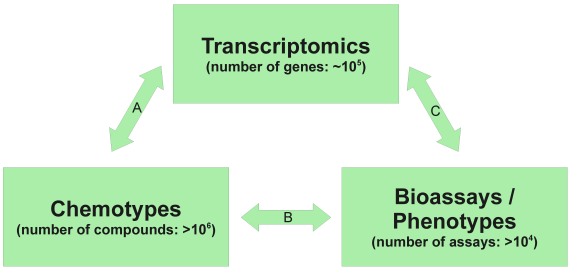

Transcriptomics
(number of genes: ~10^5)
A
C
Chemotypes
(number of compounds: >10^6)
B
Bioassays / Phenotypes
(number of assays: >10^4)

Figure 2 Available data for compound optimization.

The activity of a compound related to a (desired or undesired) target is predicted by classification or regression models (see arrow B in Fig. 2), where structural descriptions of compounds or, alternatively, similarities between compounds are used to predict the outcome of a phenotypic screen or a specific bioassay. Using these classification and regression models, new compounds can be found or existing compounds can be modified such that off-target effects are minimized while on-target effects are maximized. In practice, an objective has to be defined which trades on-target against off-target effects (it is common to speak of "Multifactorial Compound Optimization").

Biological activity including on-target and off-target effects can be measured by 'omics technologies. Again classification and regression models are constructed which now predict biological activity given by 'omics data from the structural description or similarities of compounds (see arrow A in Fig. 2).

To summarize, by interrelating chemistry, phenotype, and 'omics data, on-target and off-target effects can be predicted for a large set of candidate compounds in the compound optimization step, of which only the most promising ones need to undergo experimental validation.

All the data mentioned above (see Fig. 2) are typically high-dimensional, noisy, and technically biased, while only few samples or cell lines are available. These properties of the data entail the demand for advanced machine learning techniques and cutting-edge statistics. Data analysis starts with quality control and preprocessing. Then filtering techniques are needed to reduce dimensionality and finally structures in the data have to be identified. In a next level, these different data sources have to be combined and dependencies have to be recognized.

Summarizing, new computer science tools are required to tackle the computational challenges in early drug design. The more advanced those methods are, the more they are robust to data deficiencies, and the better the interplay between the different steps, the more helpful the results will be for early-stage drug development.

4.3 Conclusions

Combining high dimensional data from genomic technologies with chemical information of structures and classical measurements of compound effects in biological assays (e.g., biochemical or phenotypic assays) the seminar participants discussed various ways of how these data could complement established methods in computational chemistry. See Fig. 3 for topics covered in the seminar.

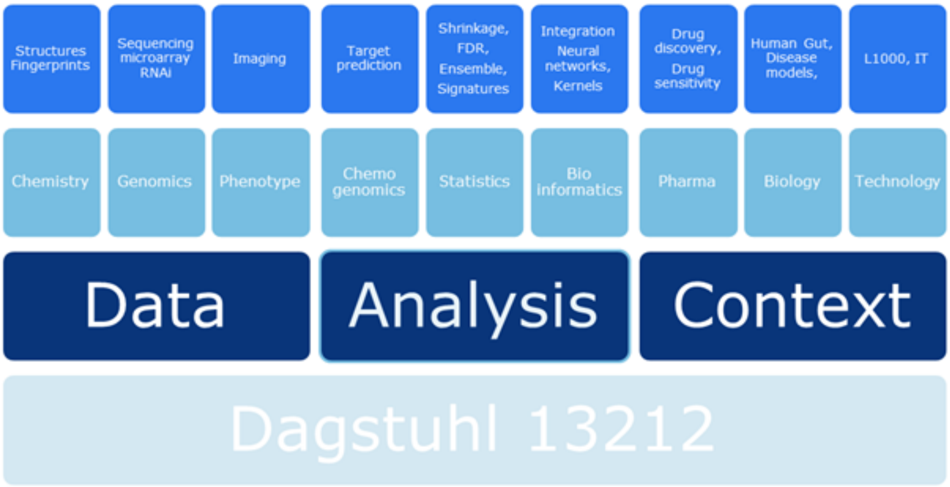


Figure 3 Summary of the topics covered in the seminar.

Being able to discover and utilize connections between the three data types has also been the focus of the QSTAR project (IWT-funded project of Janssen and academia). Some success stories of using transcriptomics in early drug design were presented by members of the QSTAR consortium. Examples of methodology being explored has been presented were transcriptomics was related to chemistry and to biological assays.

References

1 Mahrenholz C., Abfalter I., Bodenhofer U., Volkmer R., Hochreiter S.: Complex networks govern coiled coil oligomerization – Predicting and profiling by means of a machine learning approach, Mol Cell Proteomics, 2011 (doi: 10.1074/mcp.M110.004994)

2 Hochreiter S., Bodenhofer U., Heusel M., Mayr A., Mitterecker A., Kasim A., Khamiakova T., Van Sanden S., Lin D., Talloen W., Bijnens L., Göhlmann H., Shkedy Z., Clevert D.: FABIA: Factor Analysis for Bicluster Acquisition, Bioinformatics, 26(12): 1520–1527, 2010 (doi: 10.1093/bioinformatics/btq227)

3 Talloen W., Hochreiter S., Bijnens L., Kasim A., Shkedy Z., Amaratunga D., Göhlmann, H.: Filtering data from high-throughput experiments based on measurement reliability, National Academy of Sciences of the United States of America, 107(46): E173–E174, 2010 (doi: 10.1073/pnas.1010604107)

4 Hochreiter S., Clevert D., Obermayer K.: A new summarization method for affymetrix probe level data, Bioinformatics, 22(8): 943–949, 2006 (doi: 10.1093/bioinformatics/btl033)

Participants

- Dhammika Amaratunga
Johnson & Johnson, US
- Andreas Bender
University of Cambridge, GB
- Ulrich Bodenhofer
University of Linz, AT
- Chas Bountra
University of Oxford, GB
- Javier Cabrera
Rutgers Univ. – Piscataway, US
- Aakash Chavan Ravindranath
University of Cambridge, GB
- Hinrich Göhlmann
Janssen Pharmaceutica – Beerse, BE
- Jelle J. Goeman
Leiden University Medical Center, NL
- Sepp Hochreiter
University of Linz, AT
- Wolfgang Huber
EMBL Heidelberg, DE
- Murat Iskar
EMBL Heidelberg, DE
- Adetayo Kasim
Durham University, GB
- Samuel Kaski
Aalto University, FI
- Günter Klambauer
University of Linz, AT
- Leo Lahti
Wageningen University, NL
- Justin Lamb
Genometry Inc – Cambridge, US
- Johannes Mohr
TU Berlin, DE
- Gianluca Pollastri
University College Dublin, IE
- Ziv Shkedy
Hasselt Univ. – Diepenbeek, BE
- Willem Talloen
Janssen Pharmaceutica – Beerse, BE
- Oswaldo Trelles
University of Malaga, ES
- Bie Verbist
Ghent University, BE
- Jörg Kurt Wegner
Janssen Pharmaceutica – Beerse, BE